Classification of Social Media Event-Discussions Using Interaction Patterns: A Social Network Analysis Approach

by

Petrus Schoonwinkel

Thesis presented in fulfilment of the requirements for the degree of ***Master of Arts in Socio-Informatics*** *at the University of Stellenbosch*

Supervisor: Dr. L.A. Cornelissen

Co-supervisor: Dr. D.A Parry

December 2020

Stellenbosch University https://scholar.sun.ac.za

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 2020/08/27 .....

Copyright © 2020 Stellenbosch University
All rights reserved.

Stellenbosch University https://scholar.sun.ac.za

Abstract

Classification of Social Media Event-Discussions Using Interaction Patterns: A Social Network Analysis Approach

P. Schoonwinkel

Department of Information Science,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MA (Socio-Informatics)

December 2020

This thesis uses social network analysis to explore the classification of social media discussions, utilising network structure derived from interactions on Twitter, while requiring minimal domain knowledge. In academia and industry, researchers strive to understand the patterns of interaction between actors on social media platforms, and how their actions may relate to particular events, topics, network characteristics, personalities and characters, among other factors. From literature, it is found that researchers in a wide range of disciplines lack the tools to classify in a variety of event-discussions. Further exemplified with the scenario where topics of interest to researchers on social media can overlap and that users are often engaged in a multitude of topics simultaneously, an approach to classification that necessitates minimal prior domain knowledge on the contents of the datasets is required. This study is a proof of concept for the use of network metrics to characterise and classify a diverse set of events that were discussed on social media.

To classify social media data, one can utilise unsupervised machine learning methods. From the literature it is found that a multitude of clustering methods with regards to social media has been explored, in multi-media, networks, textual and other contexts. However, only limited approaches to classifying social media data—specifically Twitter—in terms of their network structure have been explored. This study does not aim to replace those methods but add to an array of tools that can be used by researchers, both in academia and in industry, to maximise the value obtained from social media data. In order to obtain metrics whereby to perform classification, a novel approach to modelling interactions with the data source, Twitter, was developed and a set of network measures and

Stellenbosch University https://scholar.sun.ac.za

data descriptors that characterise the data were explored. The network measures and data descriptors were subjected to dimensionality reduction to account for co-variance in the measurements and to evaluate the contribution of each network measure, in order to expand the literature on what they define in the context of this study. The resulting principal components were used to classify the discussions of diverse events and the quality and quantity of clusters were evaluated. Finally, a set of tests and criteria were defined with which the research question was addressed.

The study found that the approach produced an optimal number of clusters with reasonable structure quality without requiring any domain knowledge to produce them. Although the method proposed in this study is effective in finding underlying patterns and similarities, it mainly serves to point researchers in the right direction, more detailed analysis is necessary for definite conclusions and labelled categorisation. The study recognises the prior work performed in classifying social media data and recommends that future work include a wide variety of user features, sentiment, topic, and network measures. Furthermore, the study can be expanded upon by testing alternative dimensionality reduction and clustering methods at each stage of the proposed approach. The study furthered the understanding of classifying social media data in terms of social network analysis and the various network measures and data descriptors that was discussed.

Stellenbosch University https://scholar.sun.ac.za

Uittreksel

Klassifikasie van Sosialemedia Gebeurtenisbesprekings deur Interaksiepatrone: 'n Sosialenetwerkanalise Benadering

("Classification of Social Media Event-Discussions Using Interaction Patterns: A Social Network Analysis Approach")

P. Schoonwinkel

Departement Inligtingwetenskap,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MA (Sosio-Informatika)

Desember 2020

In hierdie tesis word sosiale netwerkanalise gebruik om die klassifikasie van sosiale media-besprekings, met behulp van die netwerkstruktuur afgelei van interaksies op Twitter, sonder die nodigheid van domeinkennis te ondersoek. In die akademie en die industrie poog navorsers om die interaksiepatrone tussen akteurs op sosialemediaplatforms te verstaan, en hoe hul optrede onder andere verband hou met spesifieke gebeure, onderwerpe, netwerkkenmerke, persoonlikhede en karakters. Uit die literatuur blyk dit dat navorsers in 'n wye verskeidenheid van dissiplines nie die gereedskap het om in 'n verskeidenheid van gespreksonderwerpe op sosialemedia te klassifiseer nie. Dit is 'n algemene scenario dat onderwerpe wat navorsers op sosiale media van belang is, kan oorvleuel en dat gebruikers dikwels terselfdertyd besig is met 'n menigte onderwerpe. Hierdie gevalle motiveer die benadering tot klassifikasie wat 'n minimale kennis van die domein oor die inhoud van die datastelle noodsaak. Hierdie studie is 'n konsepbewys vir die gebruik van netwerkmetings ('network metrics') om 'n uiteenlopende reeks gebeure wat op sosialemedia bespreek is, te karakteriseer en te klassifiseer.

Om gegewens op sosiale media te klassifiseer, kan u gebruik maak van masjienleermetodes wat nie onder toesig is nie. Die literatuur toon dat 'n menigte groeperingsmetodes ('clustering methods') van sosialemedia ondersoek is, in multimedia, netwerke, tekstuele en ander kontekste. Slegs 'n beperkte aantal benaderings tot die klassifisering van sosialemediadata, spesifiek Twitter, is ondersoek. Hierdie studie is nie daarop gemik om hierdie metodes te vervang nie, maar dra by tot 'n verskeidenheid instrumente wat navorsers, sowel in die akademie as in die industrie, kan gebruik om die waarde wat uit sosialemediadata

Stellenbosch University https://scholar.sun.ac.za

verkry word, te maksimeer. Om maatstawwe ('metrics') te bekom waardeur klassifikasie uitgevoer kan word, is 'n nuwe benadering tot die modellering van interaksies met die databron (Twitter) ontwikkel, en 'n stel netwerkmetings en databeskrywers ('data descriptors') wat die data kenmerk, ondersoek. Die netwerkmaatstawwe en databeskrywers is aan dimensie-vermindering onderwerp om rekening te hou met die kovariansie van die metings en om die bydrae van elke netwerkmaatstaf te evalueer. Die gevolglike hoofkomponente is gebruik om die besprekings van uiteenlopende gebeure te klassifiseer en die kwaliteit en hoeveelheid trosse ('clusters') is beoordeel. Laastens is 'n stel toetse en kriteria gedefinieer waarmee die navorsingsvraag aangespreek is.

Die studie het bevind dat die benadering 'n optimale aantal groepe met 'n redelike struktuurkwaliteit opgelewer het sonder dat enige domeinkennis nodig was om dit te produseer. Alhoewel die metode wat in hierdie studie voorgestel word, effektief is om onderliggende patrone en ooreenkomste te vind, dien dit veral om navorsers in die regte rigting te wys, maar meer gedetailleerde ontleding is nodig vir definitiewe gevolgtrekkings en benoemde kategorisering. Die studie erken die vorige werk wat gedoen is met die klassifikasie van sosialemediadata en beveel aan dat toekomstige werk 'n wye verskeidenheid gebruikersfunksies, sentimente, onderwerp en netwerkmaatreëls insluit. Verder kan die studie uitgebrei word deur alternatiewe dimensieverminderings en groeperingsmetodes in elke stadium van die voorgestelde benadering te toets. Die studie het die begrip van die klassifikasie van sosialemediadata ten opsigte van sosiale netwerkanalise en die verskillende netwerkmetings en databeskrywings bevorder.

Stellenbosch University https://scholar.sun.ac.za

Acknowledgements

I would like to extend acknowledgements to all those that proved crucial to the development of this thesis.

Eerstens wil ek vir my pa en ma uit die diepte van my hart dankie sê. Te danke aan julle menigte raad, gebed, onvoorwaardelike ondersteuning en liefde het die lang akademiese avontuur met heelwat draaie tot 'n suksesvolle einde gekom. Julle is altyd so geduldig om te luister na al my stories, so hopelik is hierdie een ook interessant genoeg. Baie lief vir julle.

To my supervisor and mentor, Aldu, who inspired me to read more and encouraged critical thinking on a wide range of topics, I say thank you. You've broadened my perspectives on more than just social network theory, although it was definitely not neglected with the number of times you had to patiently explain the same network measures to me. Thank you for always making time to talk during a very busy period.

I would also like to thank my co-supervisor, Doug, without whom this thesis would definitely not be completed in time. I'm very thankful that you always found time to provide valuable feedback and comments on my latest chapter, often only a few minutes or hours after I sent them.

Vir my broer Daniël en Irene wil ek ook baie dankie sê. Julle kuiers, saam kos maak en geduld met my ewige debatte het my gemoed hoog gehou in die besigste tyd. Ek wens julle al die beste toe met die nuwe fase van julle lewe.

To all my family and friends, I thank you for the support, encouragement and interest.

The financial assistance of the Council for Scientific and Industrial Research (CSIR) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the CSIR.

Stellenbosch University https://scholar.sun.ac.za

Contents

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

List of Figures

Stellenbosch University https://scholar.sun.ac.za

List of Tables

Stellenbosch University https://scholar.sun.ac.za

Chapter 1

Introduction

1.1 Chapter Introduction

Humans have gone from simple synchronous in-person communication to multi-layered, multi-platform complex, truly light-speed global communication as part of daily life. With a long and rich history of sociological exploration, based on their communication and interaction patterns in sociology and anthropology, researchers have only started to scratch the surface in understanding the traces people leave on these new electronic communication platforms. New electronic communication platforms have increased the complexity of how humans interact and an ever expanding set of tools have been developed to facilitate this fundamental aspect of society. Modern communication methods have seen much of these tools become digitised, with the foremost example being social networking sites, also known as social media platforms (Boyd and Ellison, 2007). However, the ability to classify individuals based on their online interactions has been brought into question. As is the case for offline interactions, there still exists a need to understand the patterns of interaction between actors on social media platforms, and how their actions may relate to particular events, topics, network characteristics, personalities and characters, among other factors. Twitter is an example of an online social networking platform where many topics that are of interest to the public domain are discussed. The popularity of Twitter, with over 350,000 posts, referred to as tweets (or retweets), per minute and 318 million monthly active users (Twitter, 2019; Sayce, 2020), has led to the necessity of understanding this form of communication. This is due to the fact that these contributions are part of the networking processes of many academic and industrial contexts, where individuals share and exchange information with other members in these online communities (Stefanidis *et al.*, 2013). Apart from networking, individuals and companies also interact on social media for other reasons, and these discussions often form along a variety of events, reflective of something that happened in the physical or digital world. For instance, Acar and Muraki (2011, p. 394) indicated how an event can be digitally captured by informa-

Stellenbosch University https://scholar.sun.ac.za

tion sharing in a study regarding Twitter usage during the March 2011 tsunami in Japan. However, Twitter is not only used to share information; its short form messaging style is also used to facilitate a wide range of discussions, of which a considerable amount is controversial and has large political impact (Popescu and Pennacchiotti, 2010, p. 1875). While disaster and political events are often relatable issues to the broader public, opinions on social networks can often also have a significant effect on organisations' decision making and policies (Jansen *et al.*, 2009, p. 2171). Examples of such organisations include political parties, marketing divisions, social media developers, human resources, disaster response teams, customer relations and event planners. There are also individuals and groups of personalities online which form new organisations, also highly dependent on the interactions and networks of which these discussions consist (Jansen *et al.*, 2009, p. 2169). Sigala and Chalkiti (2015, p. 55) highlight the need to understand and manage social networks internally as well as externally. It is therefore of value to these organisations to understand the network structure of online discussions that may impact them. Similarly, researchers in the social sciences and humanities have identified the need to incorporate and evaluate the measures that are traditionally applied in their specific disciplines, in order to understand a variety of online events that affect society, such as the aforementioned political, disaster and other events (Reale *et al.*, 2018, p. 299). These events can also overlap. Users on social networking platforms are often engaged in a multitude of topics, in many cases relating to real world events, current affairs, opinions and viewpoints. These online discussions could lead to underlying interaction patterns that have not been anticipated by researchers and may provide unique insights when processing the data.

Understanding these digital conversations around a variety of situations can offer the opportunity to learn about the type of interaction between users around such events, as well as discovering the network patterns that would allow anyone to identify them without necessarily being familiar with them. In other words, the ability to classify social media data, without prior domain knowledge will make this approach accessible to more researchers. In this context, classification refers to the action or process of classifying something according to shared qualities or characteristics (Duda *et al.*, 2001, p. 12). However, for the purpose of this thesis, an approach known as *clustering* will be applied to classify or group the social networks. This is the task of grouping a set of individuals or topics in such a way that those in the same group—called a cluster—are more similar to each other according to a specific set of criteria. In other words, the set of individuals or topics have more similar network patterns or other measures, than to those in other clusters. Very little attention has been afforded to the value of using network characteristics to support the clustering of Twitter data. This is not to suggest that clustering of social media data has not been attempted. Himelboim *et al.* (2017) simultaneously clustered and classified a wide range of discussions on Twitter relating to real world events. The study proposed concepts and methods for clustering patterns of information flow on Twitter based on network

Stellenbosch University https://scholar.sun.ac.za

structures. The study found that the approach was effective and created six typologies of Twitter topic-networks. However, Himelboim *et al.* (2017, p. 2) used only four metrics; namely *centralisation*, *density*, *modularity* and the fraction of users with no connections to characterise Twitter discussions.[1] The present study will aim to build upon these metrics and apply additional measures and data descriptors to achieve a more comprehensive understanding of the contributions of these measures and how useful they are for this analysis. Similarly, a study by Ferrara *et al.* (2013, p. 1) found that the popular communication form, memes, could be classified on Twitter—by clustering—in leveraging various content, metadata, and network features of tweets. Other studies that cluster social media data will be explored in a following section. But the use of Twitter specific network features has inspired the methodology with which this thesis will obtain the network characteristics of the discussions. Ferrara *et al.* (2013, p. 1) utilised primitive network features found in the tweet object obtained directly from Twitter, such as the `hashtag`, `mentions` and `tweet text` (see Table 3.1 in Chapter 3), to perform the clustering analysis. These features are present in every discussion on Twitter regardless of their context and could provide a method for clustering the data for any research purpose. This thesis therefore aims to explore the other options that network features and Social Network Analysis (see Section 2.3 in Chapter 2) have to offer and the benefits that this approach adds to utilising the interactions on Twitter to create networks.

1.2 Research Problem and Objective

The specific problem to be addressed in this thesis involves differentiating between the discussions that took place on Twitter regarding a variety of events using the interactions that those discussions consist of. This approach is applied with minimal to no domain knowledge, since events that are researched range from a broad spectrum of contexts, some political, some social and some disasters. The research objective is therefore to explore the clustering of events, as discussed on social media, by using the interactions to create a network whereby the various discussions can be characterised. In addition, it will discuss how much each of the network measures measures contributed to differentiating between the discussions and how useful they were in this context. This objective will be approached with minimal to no domain knowledge on the various event-discussions that are analysed in this study.[2]

[1] Centralization indicates the degree to which a few members hold the most number of connections in the network (Wasserman and Faust, 1994, p. 178); Density of a network is the proportion of possible relations that are actually present in the network (Wasserman and Faust, 1994, p. 314) and the modularity of a network determines how separated are the different actors from each other, in order to create subgroups (Gerlach *et al.*, 2018, p. 1). These first two measures will be explained in further detail in Chapter 2.

[2] Event-discussions is a term that will be used throughout this thesis to refer to discussions that were held on or revolved around a particular event.

Stellenbosch University https://scholar.sun.ac.za

1.3 Research Question

The advantage of working with social media data is that, by definition, it revolves around users interacting with one another. This provides an avenue for this study to analyse the network patterns of these interactions to obtain objective measures with which to computationally determine their characteristics. These characteristics would manifest in the form of a network structure for those events. This leads to the research questions posed for the study:

- *Can discussions regarding events on social media be successfully clustered, by utilising the network structure derived from interactions on Twitter, with minimal domain knowledge?*
- *If the discussions can be clustered, how do the network measures contribute to the clustering results and which of the measures are useful for this analysis?*

1.4 Research Approach

The nature of the problem under investigation in this study is understanding the underlying interaction patterns that are hidden in discussions online. As in the study performed by Himelboim *et al.* (2017, p. 2), the metrics used to describe these patterns and the characteristics of the network reflect how the users on Twitter interact during the discussions. By investigating all the interactions per event an understanding of the network structure is obtained. Similarly, the approach to clustering the discussions in an unsupervised method, by Himelboim *et al.* (2017, p. 2), will be expanded upon in this thesis, since manually labelling the data without prior knowledge on what the datasets contain is unfeasible. The aim of this thesis is not to categorise the events based on pre-defined labels for the datasets. Rather, it is to broaden the understanding by using different datasets, measures, and methods to provide further information on whether discussions regarding events on social media can be successfully clustered to maximise the value obtained from this research interest. The growth of social media as a platform for commercial, social and political research generates the need for methods that maximise the value obtained from gathering social media data (Cui *et al.*, 2018, p. 16). This task is also difficult due to the ambiguity in textual content typical of Twitter, where topics can overlap, and where users are often engaged in a multitude of topics. For example, in contexts where Twitter was used by first responders to gain information in disaster events, they struggled to glean actionable knowledge due to the aforementioned factors (Ashktorab *et al.*, 2014, p. 1). To address this problem, a variety of measures will need to be explored that can identify the underlying structures and patterns in the data. The approach will need to utilise methods that can be applied to a wide variety of disciplines in the

Stellenbosch University https://scholar.sun.ac.za

social media domain, must not be affected by a lack of domain knowledge and will provide objective metrics that can be cross-validated.

The first step of the research is to consider the options and methods for obtaining a variety of online interaction data that revolve around real-world events. This will be followed by an exploration of which network measures could provide a distinctive set of values with which to characterise a social media discussion. This step will require a broad approach to include as many network measures that could provide insight, since without domain knowledge it is not yet known which specific measures will have the largest impact on clustering a number of vastly different events on social media. Presuming that only some of the measures may prove effective, a method will need to be devised to determine how much each measure contributes to defining the network structure of the events. Finally, a clustering method will be applied, for which the quality of the results will need to be computationally evaluated.

In order to differentiate the digital conversations on social media from general conversations, the study requires a variety of discussions about a diverse set of events, such as disaster centred data and political discussions, as well as more homogeneous conversations around other topics such as sporting events. Each dataset will be analysed by obtaining values that describe their network structure with a number of network measures. Once the measures have been produced, the variance between the results will form the basis of the classification analysis.

1.5 Related Work

The problem of classifying social media data has been studied extensively in the database, machine learning, multi-media and textual analysis contexts. In most cases, the classification approaches utilised clustering methods to find similarities in the data and group them based on the clustering results. An extensive overview of clustering algorithms was compiled by Jain and Dubes (1988) in a variety of engineering and scientific disciplines. In this compilation, Jain and Dubes (1988, p. 1) define clustering analysis (or classification in the context of this study) as organising data by abstracting underlying structure either as sub-groups of data points or as a ranking of sub-groups. Several studies that will be explored in this section applied a wide variety of approaches to the types of data gathered, methods for gathering, clustering methods and machine learning algorithms in the context of social media. The studies that explored the classification of social media will be discussed and how their approaches may be relevant to the present study with regards to improving the results produced by this thesis.

One such relevant study by Amaro *et al.* (2016, p. 5) approached the problem of classifying social media data from a tourism perspective and utilised more traditional social science methods in their data gathering. The study obtained responses via questionnaires concerning how the participants used social media to plan recreational travels. The results obtained from the approach deter-

Stellenbosch University https://scholar.sun.ac.za

mined clustering methods to be effective and validated by outside research with regards to how social media affects travel based decisions (Zhang *et al.*, 2017, p. 21). Moreover, the methodology claimed to contribute to the tourism literature for developing measurement scales for social media use (Amaro *et al.*, 2016, 12). Zhang *et al.* (2017, p. 21) also supported the use of social media platforms, such as Twitter, for a low-cost and real-time method to capture relevant data. In a similar fashion, the present study aims to determine the utility of the network measures for the specific context in which they will be applied. The goal is to assist future research in the theoretical support for what characteristics certain network measures provide insight to.

As noted previously, Himelboim *et al.* (2017) considered classification on Twitter from the perspective of information flow. From the research objective in the study by Himelboim *et al.* (2017) namely, to use interactions or information flow that emerge into complex social network structures to cluster event-discussions is similar to the objective of the present study. Himelboim *et al.* (2017) segmented 60 Twitter topical social media network datasets into six unique patterns that share a set of characteristics, based on four network measures. The six structures of information flow were not labelled beforehand, but rather identified by utilizing a three-step classification model. This thesis will follow a similar trajectory: no pre-defined or labelled categories will be prepared before the network measures are applied to the dataset. However, as was deemed necessary in the limitations section of the paper by Himelboim *et al.* (2017), the present study will include a wider variety and a greater number of network measures to further analyse the contribution and usefulness of each measure to the study. Furthermore, data descriptors will also be used to account for any changes in network structure based on decisions made and factors that affect the data structure at the time of data collection.

Apart from social networks, there are other analysis methods that can be utilised to characterise social media data; for example, *sentiment analysis*.[3] Prior work in sentiment analysis on social media platforms, such as Twitter, has been performed on individual tweets or on user-level sentiment as claimed by Zhu *et al.* (2014, p. 2). In their study, they aimed to apply classification in a combined approach to multilaterally address sentiment found in tweets, users and features of the platform into a novel sentiment class. The study produced an online algorithm that was both efficient in running time and clustering quality (Zhu *et al.*, 2014, p. 25). The study faced challenges in labelling portions of the data for the supervised machine learning approach as well as the accuracy of the labelled data. In this thesis, the study aims to computationally evaluate the quality of the clustering results in order to determine whether the clustering methods produced high quality structures. However, the requirement of no prior domain

[3]The process of computationally identifying and categorizing opinions expressed in a piece of text, in order to determine whether the text is on average positive, negative, or neutral (Agarwal *et al.*, 2011, p. 33)

Stellenbosch University https://scholar.sun.ac.za

knowledge also necessitates the use of methods that do not require human input.

Of importance to this study is the applicability of analysing social media and comparing digital and physical spaces, in terms of how both spaces play a central role in shaping their behaviour. To address the research question—the need for classifying social media events by analysing the online discussion thereof—necessitates the existence of a network structure that is similar depending on the event. Croitoru *et al.* (2015, p. 16) explored this phenomenon and found that applying their clustering analysis to large-scale real-world events revealed the eventual alignment of the online and offline communities. Their study observed notable differences in the in- and out-degrees of offline and online communities:[4] the offline communities tended to consume more information than they were producing. This further supports the theory that participants of an event are influenced or even affected by the online discussion, regardless of whether or not they participate (Gaál *et al.*, 2015, p. 195).

The aforementioned study by Ferrara *et al.* (2013, p. 1) utilised network features of Twitter, for example `mentions` (See Figure 2.5 in Chapter 2), in their classification approach. Ferrara *et al.* (2013, p. 1) defined several similarity measures by leveraging various content, meta data and network features of tweets. The goal of this study was to combine these measures in different ways for classifying memes. The study explored several approaches that this thesis aims to utilise, including similarity measures for clustering, *k-means* and Twitter features.[5] The study proposed that pre-clustering their datasets produced improved results. This suggests that categorising the datasets obtained in this thesis might provide better results. Furthermore, computationally determining the cluster amounts and quality proved to produce more reliable k-means clusters for the study and will be used for this thesis.

The approach proposed by the present study is only the first step in classifying social media data. A multitude of clustering methods with regards to social media has been explored, in multi-media, textual and other contexts. A few of these have been listed above. This study does not aim to replace those methods but add to an array of tools that can be used by researchers, both in academia and industry, to maximise the value obtained from social media data. The network measures approach aims to explore the underlying patterns present in this data, obtaining valuable information on the network structure, and classifying the patterns of events online. Such a tool could be integrated into a dashboard where researchers provide streams of data, and the approach classifies the events, along with providing their network characteristics. Furthermore, to address the second research question, this study will also discuss the real world insights that particular network measures provide for a given dataset and how the results of this

[4] In-degree refers to the number of interactions directed at an actor, whereas out-degree refers to the number of interactions an actor directs at any other actor in the network. (Wasserman and Faust, 1994, p. 163)

[5] K-means is common method for partitioning a dataset into a distinct number K of non-overlapping clusters (James *et al.*, 2013, p. 386)

Stellenbosch University https://scholar.sun.ac.za

study may compare with the theoretical characteristics that those measures are expected to supply.

A wide range of insights and improvements to this study were found in related work and further supported the position and function of this study. The suggestions in the literature will be tested to potentially improve the results of this study, as well as to satisfy the requirement for more studies that do not depend on prior domain knowledge.

1.6 Chapter Breakdown

This section will preface the thesis by outlining each chapter, with particular attention to the objective of each chapter, while providing linking thoughts between chapters.

In Chapter 2, the options for data collection and limitations thereof will be explored. Thereafter, a theoretical background regarding Social Network Analysis (SNA) will be explored. The aim of SNA is to understand a community by mapping the relationships that connect them as a network, and then trying to draw out key individuals, groups within the network components, and associations between the individuals. The various metrics that may be used to determine the difference between the networks are explored in this chapter, and they are categorically referred to as network measures. The chapter will explain the different measures and which differences they highlight. Every possible measure will be tested for the range of impact, but the weighting of each metric will be determined in Chapter 3. Chapter 3 sets out the methodology of the analysis, from the data collection to creating the networks from the datasets, as well as how the various network measures are calculated. The approaches to producing variables and accounting for co-variates will be discussed and the clustering method and tests to answer the research questions will be explored.

Chapter 4 will explore the output of the applied methodology and actual choices performed to address the primary research question. The descriptive statistics of the data and measurements, as well as the results from the clustering method will be presented. An in-depth analysis of the findings will be performed with suggested interpretations of network measures and classification results.

In order to address the second research question, Chapter 5 will lead to an extended discussion, with reference to the findings in Chapter 4, about the implications for the network measures. This chapter will discuss how each network measure contributed to the clustering results, and which of them were useful. The limitations that impacted the research will be discussed and how future work can improve upon the results produced by this study. The chapter ends with the conclusion which draws all the chapters together in a final assessment.

Stellenbosch University https://scholar.sun.ac.za

Chapter 2

Background

2.1 Chapter Introduction

As outlined in Chapter 1, the primary research question posed in this thesis concerns the clustering of various events that were discussed on Twitter. This chapter considers possible approaches to addressing the problem. It will briefly outline the available options for the collection of Twitter data; indicate how that data could be analysed to provide insight for answering the research question; give a description of the final choice of data format; and provide an overview of the most pertinent approach to analysing such data.

2.2 Data Options for Addressing the Research Question

The research question in its simplest form asks: Is there a high degree of confidence with which discussions on Twitter can be successfully clustered by using the network structure? This dictates the source of the data available for this thesis. There are a variety of methods available for obtaining network data from Twitter. These range from traditional social science approaches to more computational approaches. In order to consider these options, one must first determine what data is made available by the platform.

Twitter is a unique social media platform that relies on users to condense the content they choose to share to 280 characters per message or *tweet*. A tweet can include photos, videos or conversations, a mention (@) directing to another user and the hashtag (#) feature that organises information into a particular topic (Rosen and Ihara, 2017; Cai *et al.*, 2015, p. 87). Another feature that adds to the idiosyncratic nature of the platform is that the vast majority of these tweets are sent with little to no privacy or invitation restrictions. In other words, the platform works as a broadcast channel that allows its users to follow other public Twitter accounts. In this way, the platform has been widely used as a form

Stellenbosch University https://scholar.sun.ac.za

of e-government and by other public sector organisation in order to communicate with the public regarding provided information and services (Alshehri and Keefe, 2018, p. 3). Twitter has also seen constant user growth up until 2018, with 330 million accounts currently active (Twitter, 2019). For these reasons, Twitter has become widespread as a context for conducting research, as a way to understand the nature and influence of interactions on social media (Cornelissen *et al.*, 2019; Alshehri and Keefe, 2018; Jansen *et al.*, 2009, p. 2186). Some examples include commercial research investigating how gambling brands use Twitter for customer interaction and marketing (Bradley and James, 2019); social contexts where a comparative analysis of the online hate speech phenomenon in multimedia content (Buscema *et al.*, 2015); and political research, for example, on whether social media could predict election results in New Zealand (Cameron *et al.*, 2016).

2.2.1 Data Sourcing

As noted in the previous section, there exist numerous examples of how research is conducted on Twitter with a wide variety of data collection methods. In the following subsection, the appropriateness of a number of popular data collection methods is briefly evaluated in relation to the research question posed in this study.

Twitter is a public platform upon which anyone can interact with one of its users, provided that their profile is public. Since the default state of a profile is public it would be possible to query users directly regarding their involvement in a discussion about an event of interest to this study. Questionnaires or surveys are the most common types of data collection methods that can be used to obtain user generated responses (Babbie, 2010, p. 270). These responses could be obtained by asking users participating in a particular discussion on Twitter to respond to a set of standardised questions. For example, who they considered to be part of this discussion, and with whom they share information (Abdullah *et al.*, 2016, p. 22). It would be possible to build the network structure from this, as is commonly done in more traditional studies using social networking analysis (Cornelissen, 2013; Frenz, 2019). While technically possible, however this method of generating data for network constructs does not scale to the number of users typically involved in a discussion on Twitter. Studies by Jansen *et al.* (2009, p. 2186) reveal that, tens of thousands of users are commonly involved in Twitter based research. Robson and McCartan (2013, p. 246) state that this traditional method of data collection is dependent on response accuracy and how representative the sample of users questioned is, since requesting a response from the entire user base involved in the discussion is near impossible. Typically, in small-scale studies, surveys to build a network are a feasible means to understand social structure of an event, organisation, or discussion (Cornelissen *et al.*, 2019, p. 7). However, as the number of individuals grows, such techniques become less tenable.

Stellenbosch University https://scholar.sun.ac.za

Another means of collecting data to produce a network structure of an event on Twitter is to interview the key people in the discussion. They would likely have the most extensive perspective on who was involved in the event on Twitter, as well as on how they interacted and what role they played. The benefits of this approach would add to the richer definition of types of relationships and sentiment between users. Although this would produce a smaller network, since the data is collected from a smaller sample of the users, it would allow for a more in–depth analysis to be performed on the data provided. But the responses from such a small sample could be biased and incomplete, without an adequate sample size to ensure variety and corroboration of responses.

As an alternative to these more traditional approaches, computational techniques provide another means of acquiring the necessary data for understanding the network structure of these events (Ledford, 2020, p. 329). Twitter provides a computational portal, an *Application Programming Interface* (API), that is particularly useful for analysing events on the platform. This API allows one to access the features of Twitter without having to use the website interface. This can be useful for actions such as posting tweets or sending directed messages with scripts in an automated process, and also obtaining large scale data sets with a high level of detail. For the purposes of this study, the API can be used to obtain the tweet object, which will be explained in a following chapter.

The tweets themselves from the discussion also have content that can be analysed. Based on the content of a tweet, there are a variety of computational methods to extrapolate who or what the tweet is referring to. For instance, *topic modelling* is a technique whereby the distribution of terms in a document is used to determine what the most prevalent topics are discussed in that document.[1] Sanandres *et al.* (2018, p. 684) applied a specific technique, known as Latent Dirichlet Allocation (Blei *et al.*, 2003, p. 1008) to Twitter data and found it illustrates the strength of topic modeling for analysing large text corpora and how it provides a way to study the narratives that people share on Twitter.

Another technique to obtain data from text is sentiment analysis, that was introduced in the previous chapter. Agarwal *et al.* (2011, p. 34) explored a wide variety of tools to perform sentiment analysis. A number of these tools are designed to take into account the various challenges that tweet text pose to sentiment analysis; namely, emoticons, hashtags, unique characters, language boundaries and more (Agarwal *et al.*, 2011, p. 33). In the same paper, Agarwal *et al.* (2011, p. 33) concluded by stating that sentiment analysis for Twitter data is not that different from sentiment analysis for other genres (of data). The research question posed in the present study requires a wide variety of methods be explored. This may add to the granularity of the network characteristic in order to, determine its effectiveness in clustering. The data obtained from analysing the content of a tweet comes with several scope and technical limitations, specifically referring

[1] Topic models are a type of statistical model for discovering hidden semantic structures in a text body, thereby extracting a list of abstract topics (Sanandres *et al.*, 2018, p. 684)

Stellenbosch University https://scholar.sun.ac.za

to what is feasible for the study with the computational capacity available. With regards to the scope, it is important to acknowledge that the textual data sources can be used as metrics for this analysis. Sentiment analysis and topic modeling can both produce metrics that do not require domain knowledge and can contribute to the ability to cluster the discussions (Goncalves *et al.*, 2013; Gerlach *et al.*, 2018). However, this requires computationally intensive analysis of the tweet text and increases exponentially with the volume of data used in the study (Cornelissen *et al.*, 2019, p. 8). It is also not related to the network structure of the discussion and is therefore regarded as not in the scope for this analysis. The lack of these issues specific to textual analysis are some the key benefits to analysing the network structure modelled from the interaction types between users, which will be discussed in detail in the next section.

2.2.2 Network Structure and Data Source Limitations

Consider the primary research question without the network structure requirement: How can the characteristics that define an event on Twitter be measured and clustered without prior domain knowledge? The research outcomes states that the ideal conclusion of this analysis would be the ability to systematically categorise datasets of discussions on Twitter without prior domain knowledge. This provides several key factors that determine the scope of this project.

First, the question requires a set of characteristics whereby the discussions can be clustered. This means that regardless of which discussion is being analysed, the metrics must be map-able across all domains. This set of characteristics will be clearly defined and universally applied in the analysis to ensure that the method does not vary depending on which datasets it is applied to. The factors used to compute the network structure are agnostic to the domain and no changes will need to be made in order to characterise events that have a variety of topics and concern multiple disciplines.

Second, other elements of a tweet or a Twitter user profile can also be considered. Media is an element that is frequently used on Twitter during discussions and offer some opportunities for different types of interactions (Shamma and Kennedy, 2010, p. 4). Media consists of any image, video or GIF that are linked in the tweet content. However, the second part of the first research question as stated above requires that no prior knowledge is necessary to understand the significance of a characteristic used to classify the discussion. Media often exist within a specific context within which it makes sense, and that is difficult to extrapolate from individual tweets, since it will require topic modelling to be performed alongside tweets that contain media. Media is also often used for different effects. For example, in political discussions news media and memes are often used interchangeably both to share information or react to the discussion. Textual communication can convey a great deal about the structure of events as usage patterns swell and recede and the textual content shifts (Shamma and Kennedy, 2010, p. 1). There are also somewhat simpler and more practical rea-

Stellenbosch University https://scholar.sun.ac.za

sons for not including media in the data sources: the analysis of images and video are incredibly computationally intensive and would require more resources and time than the scope allows. Furthermore, analysing text and media leads to ethical and privacy concerns. Although tweets that are accessible via the Twitter API are in the public domain, they may contain information that individuals did not anticipate would be analysed.

This section explored the various options for data collection and elaborated on why a network structure centric approach is utilised for this study. The following section will provide a brief background on Social Network Analysis (SNA) and graph theory that uses the structure of a network to format, analyse and visualise it's characteristics and provide the metrics that will be used.

2.3 Social Network Analysis and Network Representations

The degree to which social networks in the online world affect our daily lives has generated significant interest in quantifying and understanding these networks (Jansen *et al.*, 2009, p. 2186). SNA recognises that, like the real world, online networks also have patterns and structures to which the relationships between actors in those networks adhere (Wasserman and Faust, 1994, p. 3).[2] This is evidenced by the growing number of studies seeking to understand the effects—both positive and negative—that these services can engender for individuals, organisations and society at large (Cameron *et al.*, 2016; Bradley and James, 2019; Buscema *et al.*, 2015).

Robson and McCartan (2013, p. 3) state that researchers often have difficulty objectively quantifying the metrics with which social and behavioural phenomena are analysed. Consequently, they have to resort to triangulating various data points to mitigate the possibility of subjective perspectives. SNA makes use of the intuitive sense that the connections between the actors in any network are significant factors in determining the nature of the content discussed and what they do in that network. SNA addresses the shortcomings of traditional methods by applying graph theory to investigate and represent social structures, with a specific focus on the relationships between social actors; the patterns within these networks and the implications of these relationships on the network structure (Wasserman and Faust, 1994, p. 3). Freeman (1979, p. 216) elaborates on the empirical nature of SNA by defining the characteristics that enable this methodology. Apart from representing structures as a collection of relations between actors, SNA is also grounded in systematic empirical data, it makes extensive use of graphic representation, and is built upon mathematical and/or computational

[2]Social network analysis refers to individuals in a network as actors, while Twitter and other communication platforms refers to the individuals on their platforms as users. For the purpose of this study, the terms are used interchangeably since users are modelled as actors when referring to the underlying network patterns present in online discussions.

Stellenbosch University https://scholar.sun.ac.za

models. These models are used to represent data in various ways, which the following section will explore. One such graphical representation is a sociogram, which is a representation of the social links that a person has. In other words, it is a network representation that plots the structure of interpersonal relations in a group situation (Blake and Moreno, 1954). Three representations will be explored in the next section starting with the network representation.

2.3.1 Network Representations

To explain the various elements that SNA uses to create networks, consider the following example of 16 Florentine families. Padgett (1987), Padgett and Ansell (1989, 1993), and Breiger and Pattison (1986) have extensively analyzed this data in their discussion of local role analysis, as explained by Wasserman and Faust (1994, p. 61).

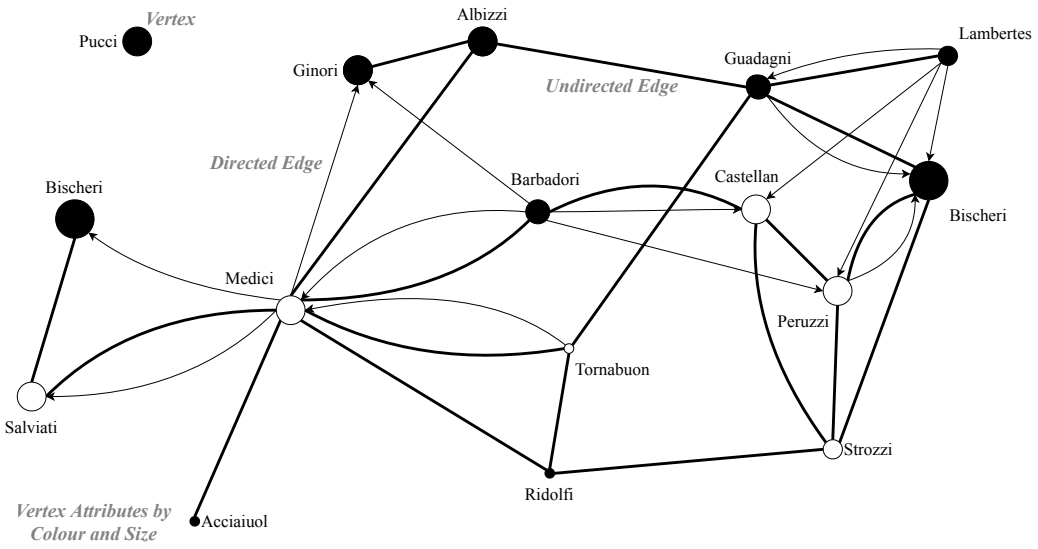


Figure 2.1: Sociogram of Florentine Relations

Note: Sociogram of marriage (grey edges) and business (black arcs) relations between 16 Florentine families (circa 1400 AD).

Figure 2.1 illustrates the graphical representation of a network as applied to the 16 Florentine families' case. In a sociogram, vertices are represented by circles, so each family is represented by a circle. In Figure 2.1, lines connecting two circles in the sociogram are known as edges and can be of two types: directed or undirected. All network representations of relations will consist of vertices $g(v)$ and edges $g(e)$, and in this way, data can be represented in various forms depending on the application

In this example, directed lines represent business relations among families. They are drawn as black arcs in Figure 2.1, pointing towards the more prosper-

Stellenbosch University https://scholar.sun.ac.za

ous family. For example, marriage relations between families are undirected lines or edges in the network. These relations are represented by grey lines in the sociogram. Furthermore, the example provides more details regarding the network, by modifying the vertices and edges. Characteristics of the social actors—for instance a person's biological sex, age, or income—can be represented by discrete or continuous attributes of the vertices in the network. In Figure 2.1, the black circles represent families who used to be members of the civic council, and the relative size of the circle represents how affluent the families were. In addition, the intensity, frequency, valence, or type of social relation can be represented by edge weights, edge values, edge signs, or edge types.

A second method of representation using the same data would be matrices. The primary matrix used in SNA is called the adjacency matrix, or sociomatrix. In graph theory, this matrix is known as an adjacency matrix because the entries in the matrix indicate whether two vertices are adjacent or not (Wasserman and Faust, 1994, p. 150). The alternative to a matrix, would be a graph that represents the relation between two sets of vertices, unequal in size (Shaham *et al.*, 2016, p. 315). This could be a set of performers and a panel of judges where the vertices represent the performers and represent the judges. If they received an exceptional score, the edge is represented by a solid line, and highlighted 1, if they received a passing score, the edge is represented by a dotted line, and 1, and no edge for a failing score.

A third common representation for relations between entities is the edge list. An edge list, also called an adjacency list, is one of the most basic and frequently used representations of a network (Golbeck, 2013, p. 13). It is a data structure in which each edge in the network is indicated by listing the pair of vertices that are connected. If the edges contain additional information—such as weight or type—it can be added with a value after the edge has been defined in the list. Ultimately, the edge list is the representation used to create the networks of Twitter data in this study. This is a matter of practicality, since only the interaction (i.e. the tweet) between the users are required to generate the network, where each interaction is an edge between users. SNA is a particularly suitable approach for predicting the category of interactions in a variety of discussions around events. The structure of the interactions contains latent information that, when measured, is likely to define the type of network. The following section will explain how the interactions on Twitter are interpreted and formatted into graphs.

2.4 Interactions on Social Media

Networks are composed of intimate, micro-level, dyadic ties, as well as ties to larger subgroups and societal organizations at the macro-level. Instead of viewing discrete actors, each with his or her unique characteristics, the world can be envisioned as composed of actors with particular relations among themselves. In the context of Twitter, all interactions between actors are indicators of their rela-

Stellenbosch University https://scholar.sun.ac.za

tionships, where the macro-level view of those interactions might form a structure unique to the event that is being discussed. In the *Handbook of Social Psychology* (Delamater, 2006), Felmlee and Faris (2006, p. 439) wrote a chapter on 'Interaction in Social Networks', which suggests that a relationship-centric approach on virtual communication brings a fundamentally structural perspective to social psychological inquiry, a view in which individuals are seen as linked to one another in a structure of ties. This structural perspective can be seen in the interaction-affordances provided by a social media platform. In the section that follows, the specific interactions afforded by Twitter will be examined.

2.4.1 Interactions on Twitter

As explained in the previous section, graph theory defines the structure with which actors and the interactions between them are represented. This section will define the various types of interactions on Twitter and situate them within this nomenclature. The next chapter will provide a detailed explanation of how Twitter interactions are represented computationally (i.e., a Tweet Object).

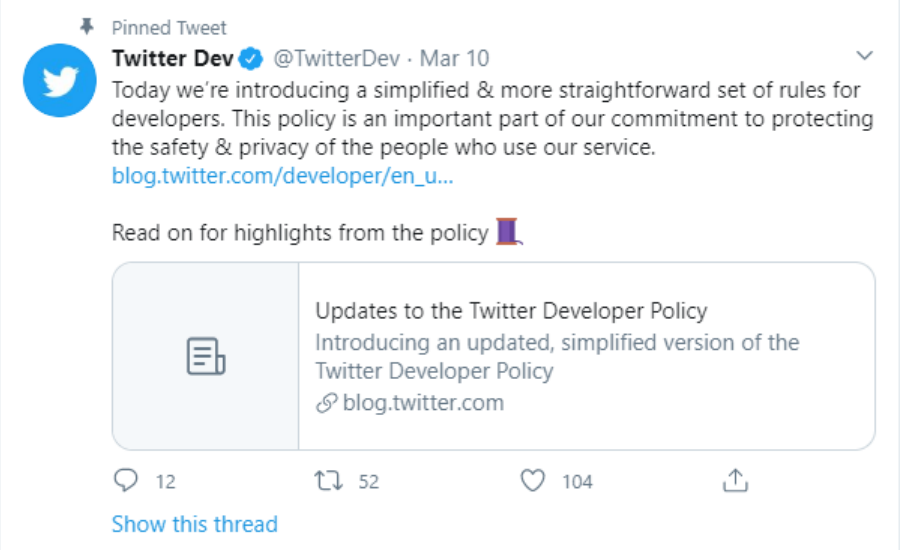


Figure 2.2: Sample Tweet

Consider the sample tweet provided in Figure 2.2, and the corresponding tweet data object. There are various elements that indicate what interaction is taking place. The user sending the tweet will always be displayed at the top of the tweet by both their `name` and `screen name`.[3] At the bottom of the tweet, it is possible to see the various counts and type of actions users took in relation

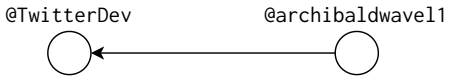
[3]Twitter differentiates between the `name` individuals chose to display and can change at any time, and the `screen name` that points a specific user in the database and cannot be modified as easily, also known as a *Twitter handle*.

Stellenbosch University https://scholar.sun.ac.za

to this tweet. Counts are displayed for the various types of interactions. This is displayed by the `comment` icon, depicted at the bottom left of the tweet, the `retweet` icon, displayed to the right of the comment icon and the `like` icon, displayed to the right of the retweet icon. Between these two elements is the content of the tweet. This example is a section of text, followed by a link and a preview of this link. The preview is only visible in the graphically rendered tweet, and is only recorded as a link in the tweet object. This tweet is not sent to any Twitter user in particular, but effectively sent to all the followers of the Twitter account `@TwitterDev`, where the `@` sign, followed by a username indicates a direct link from the sender to that account.

Figure 2.3a illustrates a direct reply sent to the example tweet depicted in Figure 2.2.[4] In this example, the Twitter interface adds a visual—'Replying to @TwitterDev'—indicating that this tweet is in reply to the sample tweet. This reply would then be one of the 12 replies indicated by the reply count in Figure 2.2. This interaction between two users is recorded in the Twitter database and can be used in the modelling of Twitter interactions in the following section.

(a) Sample Reply



(b) Directionality of Sample Reply

Figure 2.3: Twitter Reply Interaction and Model

In addition to a reply, Twitter offers further possibilities for interaction between users. These can be modelled with the same approach. The icon second from the left in Figure 2.2, represents the `retweet` icon. A *retweet* is an amplification of a tweet made by a Twitter user, an example of which is represented in Figure 2.4a.

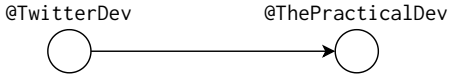
The icon and text at the top of the tweet 'Twitter Dev Retweeted' indicates that the `@TwitterDev` account chose to amplify this tweet by `@ThePracticalDev` and share it with all their followers. Where the tweet would previously have only been seen by the followers of `@ThePracticalDevs` on their feeds, it is now also seen by all of `@TwitterDevs` followers as well.

The simplest form of interaction on Twitter is known as a *mention*. This signifies a direct interaction from one user to another specific user. It is the simplest,

[4]The examples in this section will include the sample interaction as well as how they are modelled using graph theory. The following sections will discuss the modelling of interactions with Twitter data and the directionality

Stellenbosch University https://scholar.sun.ac.za

(a) Sample Retweet

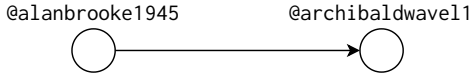


(b) Directionality of Sample Retweet

Figure 2.4: Twitter Retweet Interaction and Model

since this form of interaction does not require any prior interaction made by this or another user. For example, with a reply interaction, another user must first send a tweet for other users to reply to it. In Figure 2.5a, the tweet content contains `@achibaldwavel1`.

(a) Sample Mention



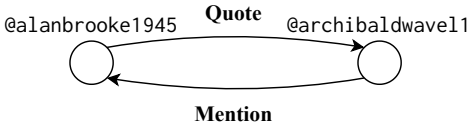
(b) Directionality of Sample Mention

Figure 2.5: Twitter Mention Interaction and Model

Stellenbosch University https://scholar.sun.ac.za

Finally, the *quote interaction* is a fourth relevant interaction form. It is possible for a user to retweet another tweet, and add their own text to the interaction. This is also known as a *retweet with comment*. This interaction would be modelled similarly to a retweet, however there is the possibility for more interactions, since the text could also mention other users.

(a) Sample Quote



(b) Directionality of Sample Quote

Figure 2.6: Twitter Quote Interaction and Model

In Figure 2.6a, `@archibaldwavel1` quote tweeted `@alanbrook1945's` tweet, in order to share it with all his followers.

To summarise, all the interactions discussed in this section are listed in Table 2.1. Using these interactions to create a network will enable the representation of all Twitter-events in a standardised format. This fulfils the requirement set out by the research question that—the same process must be used on any dataset, regardless of which discussion is being analysed. This standardisation allows the metrics to be clearly defined and universally applied in the analysis to ensure that the results are not influenced by the necessity for domain knowledge.

2.5 Modelling Interactions with Twitter Data

The four interaction types described in the preceding section are the basis of networks created on Twitter on the front-end user experience. It is possible to gain a more detailed description of the interactions, however, since these can be used in combination within a single tweet. This is due to the method Twitter uses to save the traces of their interactions. Twitter uses mentions to facilitate all interactions, with additional elements to indicate whether they are retweets, quotes or replies. In other words, an interaction on Twitter will always contain an `@username` to indicate at whom the interaction is directed. This is hidden on the front-end of Twitter for users, but visible within the data. This will be discussed further in Chapter 3.

Stellenbosch University https://scholar.sun.ac.za

Table 2.1: A Summary of Interaction Types on Twitter

Interaction	Field	Description
Mention	`user_mentions`	Array containing an object for every user mentioned
Reply	`in_reply_to_status`	If the represented tweet is a reply, this field will contain the integer representation of the original tweet's ID.
Quote	`quoted_status_id_str`	This field only surfaces when the tweet is a quote tweet. This is the string representation tweet ID of the quoted tweet.
Retweet	`retweeted_status.user.id_str`	Users can amplify the broadcast of tweets authored by other users by retweeting. This attribute contains a representation of the original tweet that was retweeted.

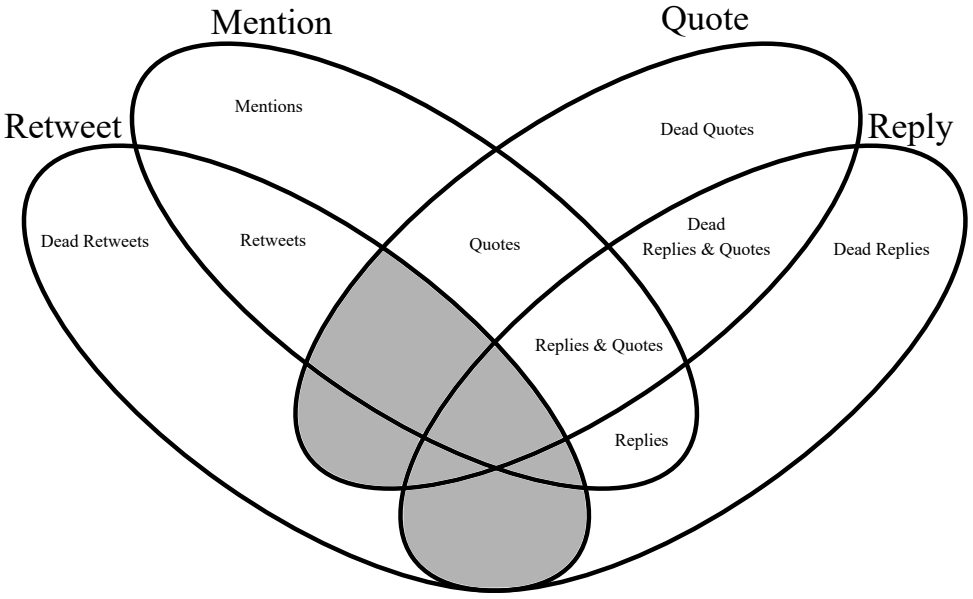


Figure 2.7: Interaction Venn Diagram

As illustrated in Figure 2.7, the intersections between Mentions and the other interactions are visible as the accurate representations of those interactions. If a tweet is recorded to not have any mention, but still has a retweet, quote, or reply attribute, that is considered a *dead* interaction. In other words, the user account on the other side of that interaction does not exist anymore. The causes for this will be elaborated upon in Chapter 3.

The depiction of the interactions in Figure 2.7 enables the identification of an additional interaction type. A *reply & quote* is created when a user adds another tweet into a reply. See Figure 2.8 for an example of this interaction. On Twitter,

Stellenbosch University https://scholar.sun.ac.za

reply & quotes are used in the user interface to reference other tweets in an ongoing discussion. This can also be used by a user to reference their own previous reply so as to create a string of coherent replies, also known as a *Thread*. This is a unique interaction that can combine separate discussions and create a new interaction pattern. This adds depth and complexity to the networks generated in this thesis. By identifying these more nuanced interactions, the granularity of types of interactions found in the datasets is increased.

Figure 2.8: Sample reply & quote.

The excluded intersections in Figure 2.7, indicated by the grey shaded areas, are interactions that cannot co-exist. If any tweets of this type are found in the data when it has been formatted, it would be an indication of a data error and should be rectified. This will help ensure both that the data is correctly interpreted and increase the quality and reliability of the data.

2.5.1 Directionality of tweets

The question of directionality is based on what the research question requires of the analysis. The question calls for characteristics that help define the network structure of an event on Twitter, that will form the basis for clustering. While it would be easier to format the interactions between users on Twitter as undirected edges, this would limit the possible number of characteristics of each network. Directed edges allow for a wider variety of measures to be applied to each dataset and increase the granularity of the analysis.

The direction of each interaction is based both on the method Twitter applies to save the type of interaction, as well as the logical order in which the interaction takes place. For example, if actor A replies to actor B, the interaction flows from A to B. Twitter uses *mentions* to indicate which other user the author of the

Stellenbosch University https://scholar.sun.ac.za

tweet is replying to, regardless of the type, as seen in Figure 2.7. Therefore, the interaction is always directed from the author of the tweet to all the interactions contained within that tweet. However, this increases in complexity when interactions are combined. Figures 2.3, 2.4, 2.5 and 2.6 contain examples of the types of interactions that can be modelled using graph theory. For example, Figure 2.3b uses the provided sample reply, but graphs the direction for that interaction. The two Twitter users are represented as vertices and the tweet interaction is the edge between them, with the direction indicated by the arrow. Note that Figure 2.6b contains a mention as well, embedded in the quote from `@alanbrook1945`. This will generate an additional directed edge between the two actors.

Figure 2.9a contains an example of a retweet, quote and multiple mentions, which create a network between four actors. The tweet is sent by `@alanbrooke1945`. This contains a single mention to `@archibaldwavel1`, who then quoted his tweet. `@archbaldwavel1's` quote contained two mentions: one back at `@alanbrooke1945`, and the other to a new actor `@hugh_downding`. Finally, the entire interaction was retweeted by `@haroldalex1945`. The directions of the interactions are maintained in the order they follow in the discussion, as shown in Figure 2.9b.

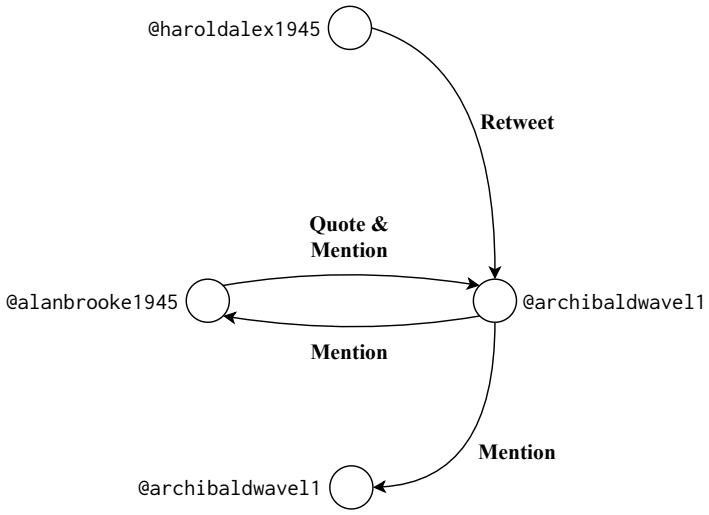
With the dataset converted into networks, the next requirement is a set of metrics to determine the characteristics of each dataset. The difference in these metrics will be used to determine how each discussion differs and whether they can be placed into recognisable categories.

2.6 Network Measures for Twitter Interactions

In the previous sections, the nature of interactions on Twitter has been outlined, with a focus on the actors and the interactions that connect them on the platform. As noted, within the sociogram representation users of the system can be represented as vertices while relations between them can be represented as edges. Individuals and groups are represented with this system and graph theory provides the methods for analysing the formal properties of the resulting sociograms (Scott, 2011, p. 25). SNA encompasses both graph theory—which is used to model the social interactions—as well as the theories underpinning the measures used to analyse the models. Theories in SNA explore the social implications of interactions with questions. For example: How does an actor's position in the network affect other actor's outcomes? How do actors affect social structure and social dynamics of the network and the overall outcomes of the network (Robins, 2015, p. 64)? An understanding of these measures and what they aim to capture needs to be explored in order to address the research question and cluster discussions on social media. SNA encapsulates a wide range of behavioural, mathematical, and social approaches to defining and quantifying the structures found in networks. Two of the seminal texts in this regard were Wasserman and Faust (1994) and Newman (2002). The authors of these books provided detailed

Stellenbosch University https://scholar.sun.ac.za

(a) Sample of a Non-Trivial Twitter Interaction



(b) Non-Trivial Twitter Interaction Graph

Figure 2.9: Non-trivial Twitter Interaction and Model

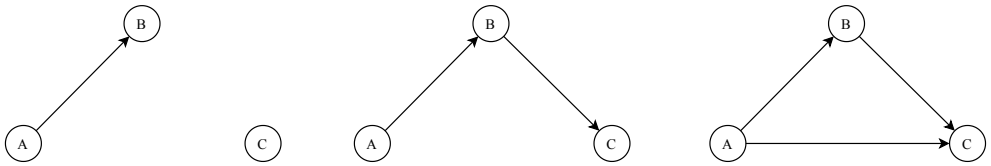
Note: Sample of a non-trivial Twitter interaction, along with the graph depicted with vertices, edges, and directions.

explanations of the measures, theories and applications that prevalent in SNA and these formalisations are expressed with relational concepts or processes. SNA has a rich tradition in theorising from a network perspective, expanding the understanding of how interactions and the measures explain the underpinning structure of a network. Also important to note, this section will briefly reference the various data descriptors—a first addition for studies in this context—and how they might apply to the various measures. The descriptors will not be explored fully in this section, since they are not network measures, but rather in detail in Chapter 3. The section which follows outlines nine network measures that are key to SNA and the goal of determining the network structure of the events.

2.6.1 Transitivity

As was discussed in Section 2.3, there are several types of network representations whereby to model social interactions. Similarly, there are four levels that

Stellenbosch University https://scholar.sun.ac.za



(a) One Link Triad (b) Two Link Triad (c) Three Link Triad

Figure 2.10: Possible network structures in a triad

Note: For transitivity, a directed network is not required to determine if a triad is transitive. However, for reciprocity, interactions between the actors must be directed.

interactions between actors can be measured. The highest level is that of the network as a whole: this contains all the actors in the discussion. The most commonly used level is a *triad*: the interaction patterns between three users. The interactions viewed between two actors is known as a dyad and a single actor could also be referred to as a node. For of this section, the example of three actors—A, B and C—will be used to visualise what each measure captures.

Figure 2.10 visualises the possible connections between the three actors. The first measure, *Transitivity*, is a property of networks that considers patterns between three actors in a network (Wasserman and Faust, 1994, p. 243). As depicted in Figure 2.10c, the relation is transitive if A interacts with B, B interacts with C and C interacts with A. Transitivity, also known as the clustering coefficient, is a graph level measurement that indicates a number of important social factors present in a network. One can evaluate the trustworthiness between two unknown participants along the social trust paths between them based on the trust transitivity properties, so to paraphrase the example above, if A trusts B and B trusts C, then A can trust C to some extent (Liu *et al.*, 2011, p. 1222). It is important to note that this measure needs not be directed, but can also be explained by simply stating that if a connection exists between A and B, then a network is transitive if a connection between A and C also exists. This is best visualised in triadic relations, where the count of the relative prevalence of the for types of relations is called *triad census*. This is also known as the *homophily* of the network, where the more transitive triples there are in a network, indicates that the actors are more likely to form connections with others similar to themselves, or where one of their connections are already adjacent to the new actor (McPherson *et al.*, 2001). This property defines the *structural balance* between triads on a sub-graph level. The term structural balance refers to groups of people and effective relations that are substantially pleasing or lack interpersonal psychological tension.

There is a range of factors that determine why people are connected in a network. Their relationships can be based on friendship, work, support, marriage, blood relation, information exchange, similar interests and more. While the reason for the connection is unknown, the measurable effect of the type of connection is visible through the structure of the network. The homophily prin-

Stellenbosch University https://scholar.sun.ac.za

ciple structures network ties of every type regarding many socio-demographic, behavioural, and intrapersonal characteristics (McPherson *et al.*, 2001, p. 429). Transitivity measures how clustered or clumped the network is. This differentiates it from the following measure (reciprocity) since it captures patterns on a triadic level. When transitivity in the network is high, it indicates the prevalence of clustering, and when it is low, the network is fragmented. By measuring transitivity, the study hopes to gain insight into the strength and characteristics by which networks are connected and thereby cluster the mean type of connection present in the network. The relevant data descriptor for this measure, and many of the following, would be how much of the original discussion was captured in the data collection. As will be explained in detail in Chapter 3, when collecting Twitter data, some data is lost over time and could lead to many triads in the original discussion missing an interaction or two. The study will investigate how correlated transitivity is to the attrition and data age of the network in the following chapters.

2.6.2 Reciprocity

As was alluded to in the previous paragraph, *Reciprocity* is a measure that is calculated on the dyadic level. Reciprocity is a measurement that originated to determine how strong the tendency is for one actor to interact with another actor, and if the second actor responds to the first (Wasserman and Faust, 1994, p. 507). This means that the measure requires the relations between these actors to be directed, and therefore quantifies the degree of mutuality. For example, consider that A and B in Figure 2.10a are interacting on Twitter. A mentions B in a tweet, thereby creating an interaction from A to B. The relation between these two actors would be reciprocated if B then replies to A's tweet. In Figure 2.10a, however, only an interaction from A to B is shown, and therefore the interaction is not reciprocated. This measure is then applied to all dyadic interactions in the network, and the tendency is represented on a scale from 0 to 1 for the entire network. If the reciprocity index equals 0, then there is no tendency to reciprocate. If it equals 1, the tendency is maximal; that is, all choices are reciprocated. It is also possible for the measure to go below 0, in which case there is a less than chance tendency for choices to be reciprocated; that is, one observes too few mutual dyads (Wasserman and Faust, 1994, p. 506).

In networks that aggregate information, reciprocity also provides a measure of the simplest feed-back process occurring in the network, in other words the tendency of an actor to respond to another in a network (Squartini *et al.*, 2013, p. 3). This principle states that vertices that have a bi-directional connection; have more social cohesion; and place more value in the information that is provided by that connection. In the case of a celebrity on Twitter who has a vast following, but who does not follow the majority back, the effectiveness with which information is shared is dependent on whether it relates to the celebrity or not. If the celebrity finds the information relevant and shares it, the interaction would

Stellenbosch University https://scholar.sun.ac.za

be reciprocated, but this is an infrequent occurrence in comparison to the amount of interactions the celebrity receives.

2.6.3 Density

Density is measured on a network level, whereas transitivity is measured on a triadic level and reciprocity on a dyadic level. The density of a network is the proportion of possible relations that are actually present in the network. This is similar to the two previous measures, but density is non-discriminative of the patterns of interactions. It is the ratio of the number of connections present to the maximum possible (Wasserman and Faust, 1994, p. 314). Whereas transitivity and reciprocity provide insight on a node level, in triads or dyads, this provides an underlying social understanding of the network. If all connections are present, then all vertices are adjacent, and the network is said to be complete. The density of a network goes from 0, if there are no connections present, to 1, if all possible connections are present. However, more insight is obtained when considering the density of sub-graphs. The measure is similar and expresses the proportion of ties that are present among a subset of the actors in a network. This is more indicative of real networks, as dense sub-graphs are indicative of cliques. This measure is used to evaluate the cohesiveness of subgroups (Wasserman and Faust, 1994, p. 315).

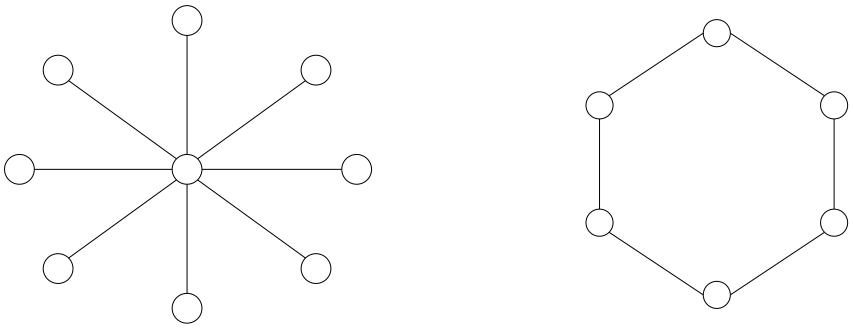
In *The Strength of Weak Ties*,[5] Granovetter (1973, p. 1362) states that networks with higher density—with more interconnected vertices—are typically connected by strong and redundant ties. This characterises a network in terms of the types of relationships present in the network. A celebrity network would mainly consist of fairly weak ties in the network as a whole, as the density of the vertices compared to the celebrity is quite low. However, combined with transitivity, one could see that the network contains lots of small clusters with strong ties (mutual interests), with the weak ties between the clusters and the celebrity. Burt (2005, p. 72) notes that networks in which people are highly interconnected are worse at transmitting information. Similarly, in social contagion scenarios, as proposed by Hodas and Lerman (2014, p. 4), the rate at which information is spread through a network has been shown to depend on its density. Because the volume of information scales with the number of friends an actor follows, highly connected actors are less likely to notice a particular piece of information.

[5] *The Strength of Weak Ties* is a seminal theoretical contribution from network theory that explores various network measures in detail and how they contribute to the understanding of the underlying patterns found in social interactions. The key contribution from Granovetter (1973) is the comparison of weak ties to bridges that allow networks to disseminate and get access to information that they might not otherwise have access to.

Stellenbosch University https://scholar.sun.ac.za

2.6.4 Centralisation

The simplest definition of actor centrality is that central actors must be the most active in the sense that they have the most ties to other actors in the network (Wasserman and Faust, 1994, p. 178). The measure of centrality can be generalised through a network that converts the measure into *centralisation*. Centrality investigates the relationships or interactions between a single actor, in this case called the *ego*, and their connections. An actor with a large degree is in direct contact or is adjacent to many other actors. This actor should then begin to be recognised by others as a major channel of relational information; indeed, a crucial cog in the network (Wasserman and Faust, 1994, p. 179). Centralisation then expresses these collectively as proportions and aims to determine what proportion of interactions are concentrated in what proportion of nodes (Wasserman and Faust, 1994, p. 177). This index reaches its maximum value of 1 when one actor interacts with all other actors, and the other actors interact only with this one, central actor. An index value of near 0 would mean that all actors in the network have the same number of relations. Figure 2.11 provides graphical representations ranging from highly centralised, Figure 2.11a to highly decentralised, Figure 2.11b.



(a) Star Network

(b) Circle Network

Figure 2.11: Sample Networks with Maximal and Minimal Centralisation

The node level insight is lost when calculating centralisation, which is a network level measure. The impact of this measure if it is highly centralised would be able to differentiate whether a few actors in a discussion are very popular or influential, but not who those actors are. In the context of online event-discussions, this is still valuable. For example, in political elections the discussion would revolve around a few key players, and centralisation would be able to identify that network structure. The key data descriptors related to centralisation, and the other centrality-based measures would be the size of the network, represented by the amount of nodes (actors) and edges (interactions) in the network. It is expected that the centrality-based measures would correlate with network size,

Stellenbosch University https://scholar.sun.ac.za

since centrality would be affected with more interactions overall and inversely affected with more actors in the dataset.

2.6.5 Assortativity

Assortativity aims to indicate if the actors in a network have a preference to connect with other actors who are similar, based on a similarity measure. Although there are multiple metrics that can be applied to similarity, in this context, most network theorists use the actor's degrees (Newman, 2002, p. 1). For example, a preference for high-degree actors to attach to other high-degree actors would result in a high assortativity. Newman (2002, p. 2) suggests that the addition of this characteristic to network models leads to a more accurate representation of the behaviours found in real work networks. In social networks, the correlation between actors with a similar degree is often found across a variety of datasets, which accords with accepted wisdom within the sociological community. Celebrities are more often connected with other celebrities of a similar social stature, as they tend to seek their equals. At first glance, this measure may seem similar to transitivity, in that they try to measure the propensity of actors in a network to interact. However, assortativity adds the additional insight of the similarity measure, in this case degree, that allows the preference to interact with other users to be equated to an increase in another network measure.

2.6.6 Closeness Centrality

Where the previous measures were applied on a network, triadic, dyadic or individual level, the remaining network measures are calculated by investigating the geodesics in a network: any interaction path between two actors in a network. *Closeness centrality* indicates how close an actor is to all other actors in the network. It is calculated as the average of the shortest path length from the actor to every other actor in the network (Golbeck, 2013, p. 27). A low closeness centrality score means that an actor is either directly connected or very few steps away from most others in the network. In contrast, outliers in the network may have high closeness centrality scores, indicating the high number of steps they need to take to connect to distant others in the network. This topological property highlights the importance of how spread out the network is, or how many interactions each actor is from any other in the network (Wasserman and Faust, 1994, p. 183). In the case of Twitter, the closer any given user is to another, the less they have to rely on other users for relaying the information from the source.

Figure 2.12 provides a graph representation of a primitive network. In this example, actor B would have a highest closeness centrality and actor A would have the lowest. A geodesic could be visualised as the interaction path from A to C, and the measures of all geodesics in a network is extrapolated to a network average to calculate the closeness centrality.

Stellenbosch University https://scholar.sun.ac.za

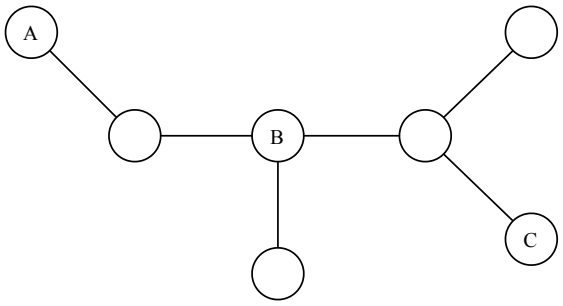


Figure 2.12: Sample Network highlighting Closeness Centrality

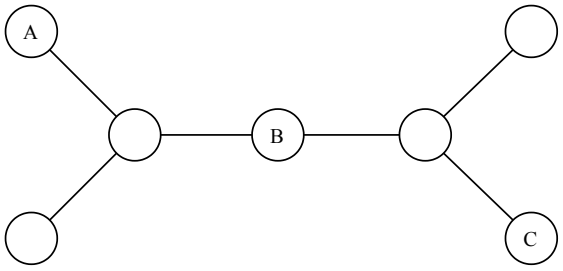


Figure 2.13: Sample Network highlighting Betweenness Centrality

2.6.7 Betweenness Centrality

Betweenness Centrality is based on the proposition that the relations between two specific actors may depend on the other actors in the network, especially the actors who lie on the paths between the two (Wasserman and Faust, 1994, p. 190). The actors between the observed actors exert a measure of control over the relation between these actors, and betweenness centrality attempts to quantify this control. In the same capacity as to which degree centrality is determined on an actor level, betweenness centrality must first be defined in the same capacity. Shimbel (1953, p. 501) and Pitts (1978, p. 286) use a count of all of the minimum paths which pass through the actor in between as a measure of the stress which that actor must undergo during the activity of the network. They also use an index between 0 and 1. The minimum value occurs when all actors have exactly the same actor betweenness index, and the maximum value has a single actor connecting all the others in the network (Wasserman and Faust, 1994, p. 190). In Figure 2.13, actor B is found to have the highest betweenness centrality, whereas actors A and C would have similarly low betweenness centrality.

Betweenness centrality measures how important an actor is to the shortest paths through the network. An actor with high betweenness may be followed by many others who do not follow the same people as the actor. This would indicate that the actor is well-followed. Alternatively, the actor may have fewer followers, but connect them to many accounts that are otherwise distant. This would indicate that the actor is a reader of many people (Golbeck, 2013, p. 30).

Stellenbosch University https://scholar.sun.ac.za

2.6.8 Diameter & Eccentricity

The *diameter* of a network is the length of the longest possible set of links to get from one actor to another in that network, that is: the longest possible geodesic. This one of the metrics that determines the various aspects of the size of a network. The diameter is a representative value of the linear size of a network. It provides insight into how sparse the network can be, though more primitive than closeness centrality. These measures are related to closeness and betweenness centrality, but serve to provide insight on a maximum and minimum level, rather than a propensity or average sparsity of the network. A high average diameter for a network could indicate that the discussion of a particular event generally occurred between smaller subgroups that are loosely connected.

The *eccentricity* of an actor is its shortest path distance from the farthest other actor in the network (Harary, 1994, p. 35)—the shortest possible geodesic. This measure only applies to directed networks, since the distance is calculated by the directed path travelled to each actor. Eccentricity is very similar to the diameter of the network, but in this application it sets out to find the highest value in the network. The reason for this is that several measures of centrality—such as the center and the centroid of a network—are based on the eccentricity of the vertices (Wasserman and Faust, 1994, p. 111). The aim for this measure is to find any correlation between the maximum distance in a network, and the centrality measures. A data descriptor that could affect the structure of the geodesics could be the length of the time span over which the data was collected. A possible correlation could exist between how long or how short the geodesics are, based on how complete the dataset was with reference to the original discussion. If only half of the discussion was captured, due to too small a window of data collection, the interactions in the discussion could be cut short. This would be evident in a correlation between diameter, eccentricity and data capture range. To summarise, all the network measures discussed in this section are listed in Table 2.2, along with their respective definitions and purposes.

2.7 Chapter Summary

In this chapter, the options for data collection and the limitations thereof were explored. Different data sources were compared based on the types of analysis that could be performed on the datasets. Thereafter, a theoretical background regarding Social Network Analysis (SNA) was discussed. The aim of SNA is to understand a community by mapping the relationships that connect them as a network, and then trying to draw out key individuals; groups within the network components; and associations between the individuals. The types of interaction on Twitter were explored. A description on network and actor levels was provided for each, as well as how those interactions would be modelled on Twitter data. The various metrics that may be used to determine the difference between

Stellenbosch University https://scholar.sun.ac.za

Table 2.2: Summary of Network Measures

Measure	**Definition**	**Purpose**
Transitivity	A property that considers patterns of triples of actors in a network. A relation is transitive if every time that actor A is in relation to B, and B is in relation to C, then A is in relation to C.	Determines the strength and characteristic by which triads in a network are connected.
Reciprocity	A measurement that originated from the need to determine how strong the tendency is for one actor to interact with another, if the second actor interacts the first.	Provides a measure of the simplest feed-back process occurring in the network, in other words the tendency of an actor to respond to another in a network.
Degree Centrality	Determines how central actors are by how they are the most active in the sense that they have the most ties to other actors in the network.	Vertices with the most social connections can be indicative of the discussion topic. These actors are the most popular actors in the network, and either led the discussion or were the most discussed.
Density	The proportion of possible relations that are present in the network.	This characterises a network by utilising the types of relationships present in the network.
Betweeness Centrality	Based on the hypothesis that the relations between two specific actors may depend on the other actors in the network, especially the actors who lie on the paths between the two.	Measures how important an actor is to the shortest paths through the network.
Closeness Centrality	Average of the shortest path length from the actor to every other actor in the network	Highlights the importance of how quickly actors can interact with any other actor in the network
Assortativity	Calculates homophily. Indicates if the actors in a network have a preference to connect with other actors who are similar, based on a similarity measure.	Leads to a more accurate representation of the behaviours found in real work networks
Diameter	Length of the longest possible set of links to get from one actor to another in that network.	Representative value of the linear size of a network
Eccentricity	Shortest path distance from the farthest other actor in the network	Find any correlation between the maximum distance in a network, and the centrality measures.

Stellenbosch University https://scholar.sun.ac.za

the networks were also explored. These are categorically referred to as network measures. This chapter explored the theory behind the various measures, methods and data sources that will be further discussed in the methodology.

Stellenbosch University https://scholar.sun.ac.za

Chapter 3

Methodology

3.1 Chapter Introduction

The previous chapter explored the concepts of Social Network Analysis (SNA) and the representation of Twitter interactions through the field of graph theory. In particular, relevant network measures in SNA were described, as well as which methods of data sourcing would be most applicable for the scope of the thesis.

The primary research question for the present study concerns the clustering of events on Twitter without prior knowledge of their content or actors (i.e., just involving characteristics of their interactions). To address this question, building on a theoretical basis from graph theory and SNA, the methodology will rely on quantitative methods. In this investigation a sample of relevant datasets reflecting the discussions on Twitter for both local and global events will be collected. For each of these datasets the nine key network measures reviewed in Chapter 2 will be calculated. While it is expected that these measures will provide useful reflections of the network structure of the Twitter discussion, it is also anticipated that, given their definitions, a degree of overlap will exist. For this reason, five co-variates reflecting aspects of the data collection process will be included for each dataset in addition to these nine measures. Furthermore, after analysing the independence of the measures, a process of dimension reduction will be performed to identify latent components that capture the variation in the data. These components will then be used as features to attempt to cluster the data successfully. This requires a set of criteria for clustering that will determine the level of confidence in this approach. These criteria will be set out in detail in the final section of this chapter.

This methodology is outlined in four sections. *First*, to address the research question, data needs to be collected. The data gathering procedure is outlined in Section 3.2. *Second*, these datasets produce a table of interactions, and network measures such as centrality, require a network structure. The datasets, therefore, need to be processed. The procedure is explained in Section 3.3. *Third*, Section 3.4 highlights the social network measures of interest, and how they will be

Stellenbosch University https://scholar.sun.ac.za

applied to the formatted networks. *Finally* dimension reduction will be explored in Section 3.5.2 as an approach to account for high correlation between the measurements, as well as the clustering methods and tests that will determine if the research question has been satisfied.

3.2 Data Collection and Rehydration

This study aims to address the question of clustering of real-world data. A variety of datasets is required, each regarding a different discussion or event that took place. It would be beneficial for the datasets to be in a broad range of categories, but also realistic relative to the discussions often found and tracked on Twitter.

Two sources of datasets were utilised for this study. The first was locally sourced by the author of this thesis and his research associates. It was initially utilised in a study on *Cross-Sample Community Detection and Sentiment Analysis on South African Twitter* (Cornelissen *et al.*, 2019). The data concerned the persistent communities on South African Twitter across 24 datasets that have been collected for a variety of events since 2014. The second source of data is from a longitudinal assessment of the persistence of Twitter datasets by Zubiaga (2018), where Zubiaga kindly made the 30 datasets used in the analysis available for academic use. In their study, Zubiaga (2018) aimed to quantify the impact of content deletion on the replicability of datasets in the long term. In order to explore this impact, their study collected a wide variety of datasets which are therefore of particular value to this study. The study contained datasets that were collected between 2012 and 2016, on a wide variety of real-world events, ranging from political to sporting to natural disasters and more. Due to data sharing limitations, however, these datasets only contain the unique numerical identifiers of each tweet. The identifiers thus needed to go through a process known as *rehydration* to obtain the full dataset containing all the necessary information from Twitter directly.

3.2.1 Rehydration

Due to Twitter's terms of service limitations it is not permitted to redistribute tweets, but only tweet identifiers (IDs).[1] Users must use the Twitter Application Programming Interface (API) to rehydrate the tweets using the IDs. The unique ID for a tweet is used to retrieve the remaining fields associated with that ID. If the tweet has not been removed, the data can be retrieved. But if the tweet has been removed, the data cannot be returned. This reconstruction process is called rehydration. Furthermore, Twitter does not allow applications to support unlimited programmatic access to Twitter data. Rather, tweets can only be accessed at a predetermined rate. This rate is determined by the level of access a user requests or pays for from Twitter. For this study, the free tier of access was

[1] Available at developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

Stellenbosch University https://scholar.sun.ac.za

utilised to access a number of 100 tweets per request, at a limit of 900 requests per 15-minute window.

The datasets for this study were rehydrated via the Twitter API using these unique identifiers to obtain all interactions present during a discussion. Apart from the Twitter terms of service, there are a number of additional considerations for rehydrating data. First, specific fields are required to build the networks. By obtaining the data from the original source (i.e. Twitter itself), the study was able to regulate strictly the quality and format of the responses. Second, when rehydrating Twitter datasets, the response contains only the tweets and users that Twitter has in its database at that time. Users or tweets that have been deleted since the tweet was sent or are user created will not appear in the dataset. This allows for additional insight into the rate of attrition of various events, by measuring the amount of users that have been deleted. A dataset with a substantial amount of users deleted could indicate that a botnet participated in the discussion and attempted to influence the outcome (Morstatter *et al.*, 2016, p. 3).

3.2.2 Obtaining Tweet Object Data with the Twitter API

The Twitter API allows for interaction with Twitter services without the need for a graphical web interface. While the Twitter API enables interactions—such as the automated posting of tweets or direct messages—for the purposes of this study, it is most useful for retrieving large scale datasets containing the relevant tweet objects. This process requires the researcher to create a free Twitter user account with which to access the platform's full features. With this account, the application process for Twitter API access begins. Since July 2018, to reduce the number of malicious users on the platform, Twitter has required individuals to apply and be manually approved for access to the Twitter API. Once access has been granted, the Twitter API also limits the number of tweets one is able to request from the platform at a given period of time. This method is the most applicable option for data sourcing, since it ensures reliable and uniform collection of a wide variety of qualitative and quantitative data. This process can also be automated, which reduces the workload and increases data gathering efficiency.

The Tweet Object All Twitter API calls return tweets that provide data encoded using JavaScript Object Notation (JSON). JSON is based on key-value pairs, with named attributes and associated values. These attributes and their statuses are used to describe objects. The tweet object has a long list of `root-level` attributes, including fundamental attributes such as `id`, `created_at`, and `text`. Tweet objects are also the 'parent' object to several child objects. Tweet child objects include `user`, `entities`, and `extended_entities`. An example tweet object is returned from the Twitter API as depicted in Listing 3.1.

Stellenbosch University https://scholar.sun.ac.za

Listing 3.1: Tweet Object in JSON format

```
{
 "created_at": "Wed Apr 20 11:01:24 +0000 2020",
 "id": 12345678910,
 "id_str": "12345678910",
 "text": "Sample Text",
 "in_reply_to_status_id_str": ,
 "in_reply_to_user_id_str": ,
 "quoted_status_id_str": ,
 "is_quote_status": ,
 "quoted_status.user.id_str": ,
 "retweeted_status.user.id_str": ,
 "user": {},
 "entities": {
    "user_mentions":[{
    "id_str": ""}]
 },
 "lang": "en",
}
```

Listing 3.1 omits irrelevant information returned by the Twitter API in the interest of brevity, but contains all the important elements for Twitter interactions in the correct format. To summarise, all of the elements of the tweet object discussed in this section are listed in Table 3.1.

Stellenbosch University https://scholar.sun.ac.za

Table 3.1: Twitter Nomenclature and Tweet Object Elements.

Twitter Nomenclature	Definition	Purpose
Tweet Object	Tweets are the basic atomic building blocks of all things Twitter. JSON object returned by the Twitter API	Contains all the information and attributes for a given tweet ID
Rate-limit	The rate limit is imposed by Twitter. It i related to the rate that data is pulled into the app.	Reduces the rate at which requests can be made to the Twitter API. Depends on level of API access.
`id`	Unique identifier of each individual tweet.	Allows API user to request specific tweets from Twitter, given the id parameter.
`user.id_str`	Unique identifier of each individual Twitter User.	Refers to specific user on Twitter, allows for the creation of networks from data.
`created_at`	UTC time when this tweet was created.	Tweet time stamp
`in_reply_to_user_ID`	If the represented tweet is a reply, this field will contain the string representation of the original tweet's author ID.	Returns the ID of the user to whom the author is replying.
`quoted_status.user_ID`	This field only surfaces when the tweet is a quote tweet. This is the string representation User ID of the quoted User	Returns the ID of the user who the author is quoting.
`retweeted_status.user_ID`	This field only surfaces when the tweet is a retweet. This is the string representation User ID of the retweeted User	Returns the ID of the user who the author is retweeting.
`user_mention_ID`	Represents other Twitter users mentioned in the text of the tweet.	An Array, can contain multiple IDs of the users mentioned by the author.
`lang`	When present, indicates a BCP 47 language identifier corresponding to the machine-detected language of the tweet text, or und if no language could be detected	Allows for identification of language used.

3.3 Data Formatting

With the tweet object for every tweet in the datasets gathered, the formatting process structured the applicable fields into interactions that can be used to create the networks. This section will detail how specific fields in the tweet object are used to find the different interaction types set out in the previous chapter and how the interactions are turned into networks.

Stellenbosch University https://scholar.sun.ac.za

3.3.1 Formatting Twitter Interaction Fields into a Network

When using Twitter data to create a network based on the interactions between users, the nodes or vertices in the network are used to represent the users and the interactions between them are the edges that connect the nodes. However, Twitter data has multiple fields in the tweet object that indicate that an interaction has taken place. In Section 2.5, it was noted that some of these fields can overlap. It is thus critical to format the interactions correctly to avoid duplicates and false interactions so as not to create an inaccurate representation of the discussion. These factors directly impact the accuracy of the measures that will be applied on the various network datasets, as will be explained later in this chapter.

The first step is to identify the fields in the tweet object that indicate that an interaction has taken place. There are four fields that, when a value is present, directly link the user that sent the tweet to the user that received it. Each field is also indicative of the type of interaction that took place. If there is a user ID in one or more of these four fields, then that interaction type is present:

- `in_reply_to_user_id_str` indicates a Reply type interaction
- `quoted_status.user.id_str` indicates a Quote type interaction
- `retweeted_status.user.id_str` indicates a Retweet type interaction
- `user_mention_id_str` indicates a Mention type interaction

3.3.2 Applying Set Theory

The various types of interactions can overlap as explained in Section 2.5. For example, this would be found in the data as the same tweet having both a mention and a reply type. The main tweet types can be handled as sets on which set operations can be performed. Twitter returns boolean attributes of tweet objects indicating whether it is a quote, a reply or a retweet. The Twitter API also returns all the accounts that were mentioned in the tweet. With these four sets, the number of tweets within each set can be calculated, as well as the intersections between the sets. For instance, an individual can reply to an existing tweet, but also add an extra mention of a third party to the tweet. Such a tweet is then observed in the intersection between mention and reply (Cornelissen *et al.*, 2019, p. 2). Each tweet can be labelled according to which intersection it occupies and can therefore be assigned its specific interaction type. These edge attributes are utilised when cleaning the data. As explained by set theory, there are potential duplicates in the data, and by assigning these attributes, it is possible to safely merge edges in the database into the correct types. This method ensures a higher level of confidence that the tweets are correctly labelled and that the interactions will accurately represent the conversations on Twitter.

Stellenbosch University https://scholar.sun.ac.za

3.3.3 Creating Graphs from Twitter Networks

In Section 2.3.1, various formats in which networks can be represented were explained. At this stage in the data processing, the data has been formatted to contain the `tweet ID`, the `User ID` of the user that sent the tweet, the `User ID` of the user to whom the tweet was directed, as well as the unique type of that tweet. The next step involves producing the edge, or adjacency list. This representation requires a pair of nodes that, when listed together, indicates—in this case—a directed edge between those nodes. The structure of the Twitter interaction data perfectly fits this description. The data therefore is formatted into this representation. The edge list also allows for attributes to be assigned to an edge. For this purpose the `tweet ID`, type, `created_at`, language and from which dataset the tweet originates is added. A node-list is created by extracting all the User IDs in the edge list to be used as names for the nodes in the graph. Finally, a graph is created for each dataset by filtering the edges for each dataset and using the node-list as vertices for the graph. With the network graphs created, it is now possible to apply the various measurements listed in Section 2.6 to obtain a set of values that define the structure of each graph.

3.4 Calculating Network Measures and Data Descriptors

The next step in the analysis involved calculating the network measures that define the structure of the network made from each dataset. The description and contribution for each measure was explored in Chapter 2. The following section will explore the methods available for applying the measures on each dataset. A set of data descriptors will also be explored that will account for any variance in the network measure based on data characteristics rather than network structure.

3.4.1 Network Measures

The calculations for the network measures were performed in the R programming language (R Core Team, 2020). The functions are all from the Igraph package, compiled by Csardi and Tamas (2006).

Transitivity The overall probability for a network to have interconnected adjacent nodes is known as *transitivity*: where the higher the probability, the more tightly connected the nodes in the network are (Wasserman and Faust, 1994, p. 165). As explained in Chapter 2, transitivity is calculated on a graph level, and the average of that probability indicates the transitivity of the network. For example, if the neighbours of the actor have more connections, the probability for the network to have interconnected adjacent nodes increases. In other words, this is

Stellenbosch University https://scholar.sun.ac.za

simply the ratio of possible triads and the triads with connections between three actors in the graph. Although this measure can be calculated with directed edges, as shown in Chapter 2, for the purpose of this thesis it was calculated as an undirected graph measure, so all interactions regardless of direction were included in this calculation. The study recognises this as a limitation of the study. The justification for this choice is technical in nature, since the directed implementation is quite computationally intensive. The value produced by this calculation depends on the number of triads in the network, which is related to the number of nodes in the network.

Reciprocity Where transitivity considers the undirected connections between the neighbours of a node, *reciprocity* considers whether the ties between nodes in a network are mutual or one-directional. Traditionally, this measure was calculated by determining the ratio of the number of bi-directional connections a node has to the total connections of that node (Wasserman and Faust, 1994, p. 514). This measure produces a numeric scalar between zero and one. This is a useful indicator of the degree of mutuality and reciprocal exchange in a network, which relates to social cohesion. For example, in a graph where the ego is a celebrity, the ego may have a very large amount of total connections. But if that celebrity follows relatively few people back, the graph will return a value closer to zero when calculating the reciprocity. On the ego level, these are called dyadic relation ships, and the range of values between zero and one will depend on the overall possible directed dyads that could exist in the network. As discussed in Chapter 2, the directionality of the Twitter interactions will greatly impact this measure.

Centralisation The graph-level centrality score—also known as *centralisation*—can be normalized by dividing by the maximum theoretical score for a graph with the same number of vertices, using the same parameters (Csardi and Tamas, 2006). In other words, the number of neighbours an actor has is important. If the network is spread out, then there should be low centralisation; if the centralisation is high, then vertices with large degrees should dominate the network (Wasserman and Faust, 1994, p. 170). Given that the data in this study is directed, the centralisation function would produce a signed value for each dataset. Take the following example: if the highest-degree node in a network has 20 edges, a centralisation score of 0.5 would translate to a node with ten edges. A node with a degree of two would have a value of 0.1 for their degree centrality. Consider this example: two nodes labelled, i and j, have the same high degree (i.e., many social connections, nine in this case), but the two roles they play are very different. Node i is very central to the network and node j is on the periphery. These show that while degree centrality accurately tells us who has a lot of social connections, it does not necessarily show who is in the center of the network (Golbeck, 2013).

Stellenbosch University https://scholar.sun.ac.za

This scenario will be accounted for by including betweenness centrality in the study and will be described shortly.

Closeness Centrality The *closeness centrality* is calculated by the length of the shortest path from every actor to another in the network, per actor, averaged by the number of actors in the dataset (Freeman, 1979, p. 225). Igraph (Csardi and Tamas, 2006) applies the method supplied by (Freeman, 1979) to the network as a whole and produces a graph level index for the closeness centrality.

Density The *density* is a measure of the proportion of possible ties which are actualised among the members of a network (Wasserman and Faust, 1994, p. 101). The extent to which a network is densely interconnected impacts the rate of information flow within it. Interaction between individuals leads to shared knowledge, and shared knowledge leads to even more interaction (Wasserman and Faust, 1994, p. 102). Density is also calculated by the number of potential connections in a network, divided by the number of actual edges, resulting in a fractional value between 0 and 1. In other words, if all the possible connections in a network are saturated, it would have a density of 1. The opposite would be a network of completely isolated nodes. But since this defies the definition of network, a 0 result would indicate a lack of a network. The calculation is slightly different for directed and undirected networks, as directed networks have twice as many possible edges.

Assortativity *Assortativity* is calculated by variance in the sum of all the normalised degree distributions for each node (Newman, 2002, p. 1). However, the normalised distribution in this case is calculated with the remaining degree of any given node; that is, the number of edges leaving the vertex other than the initial edge. The Pearson correlation coefficient is used in the Igraph package to calculate the assortativity degree for the network as a whole (Csardi and Tamas, 2006). For example, if the Pearson coefficient is high, that means that connected vertices tend to have the same labels or similar assigned values. Moreover, since the assortativity function uses the directed network to calculate the degrees, this could also result in a signed return value from -1 to 1.

Betweenness Centrality The *betweenness centrality* is a probability of which actor has the highest number of relation-links; that is, a relation passing through that actor between any other two actors (Freeman, 1979). To compute betweenness, a pair of other actors, i and j, are selected and the shortest paths between those actors, also known as a geodesic (g), are found. Then the fraction of those shortest paths that include actor A are calculated. However, this is only true as long as $i \neq j \neq k$. If there were five shortest paths between the other actors, and three of them went through actor A, then the fraction would be those three paths divided by the total 5 paths, producing a 0.6 probability value (Golbeck, 2013).

Stellenbosch University https://scholar.sun.ac.za

Diameter The *diameter* is calculated by using a breadth-first search like method. Once the distances between all nodes have been found, the function returns the length of the longest geodesic for that network (Harary, 1994, 14). The formula to calculate the diameter could simply be denoted as: $max_{u,v}d$, where u and v are two any graph vertices, and $d(_{u,v})$ is a graph distance.

Eccentricity The *eccentricity* of a node is calculated by measuring the shortest directional distance from the node to all other nodes in the graph and taking the maximum. The maximum among all the nodes is then returned as the highest eccentricity for that network. The eccentricity $\epsilon(v)$ of a graph vertex v in a connected graph G is the maximum graph distance between v and any other vertex u of G (Harary, 1994, 35).

3.4.2 Data Descriptors

The network measures listed above will hopefully give substantial enough insight into the structure of the network to determine whether the networks can be clustered. The results for each measure may be skewed, however, by a number of characteristics of the dataset. A series of descriptors for each dataset will potentially indicate any variations in the data that was not caused by network structure, but rather by the data collection or search parameters. Some of the descriptors will correlate with the network measures if the measure has been correctly calculated and will increase the confidence in the reliability of the data.

Data Capture Range A field in the tweet object indicates when the tweet was sent. By subtracting the most recent date in the dataset from the oldest, the exact time frame in which the data was captured, can be calculated. This metric will enable any relationship between the *capture range* and the network measures calculated to be determined. For example, this metric will indicate whether a dataset that was collected over several weeks has inherently different network characteristics than one collected in a few days, or any impact between datasets with a similar capture range. The `created_at` field in the tweet object returns a time stamp in a format determined by Twitter. This value can be converted to a type known as Unix or POSIX time, which represents the number of seconds since the 1st January, 1970. The difference between the minimum and maximum time is calculated, and the result is used as the capture range.

Node & Edge Count The *node and edge count* are indicators of network size. The node count is the number of unique actors in a given network. This is an indication of how many actors were involved in the discussion and is a measure of network size. This measure should theoretically correlate with the centrality measures, where a variance in node count is found to produce a similar variance in the centrality measures. The node count is calculated for each network

Stellenbosch University https://scholar.sun.ac.za

by counting the intersect of the unique nodes in the nodelist and edgelist. This extra step is taken to ensure that all actors in the network are accounted for, and that the nodelist is complete. The edge count is the number of interactions in a given network. This is an indication of how much an event was discussed by the actors in the network. This is the other measure that determines the size of the network and will also correlate with centrality measures. The edge count is performed by the gsize function in the Igraph package (Csardi and Tamas, 2006), which simply counts the number of edges for a graph object and returns them as numeric value.

Data Age The *data age* is a count of the number of days between the average time stamped date of the dataset and a fixed point in present time. Twitter has applied various changes to the data structure of their database over the past few years, as well as the ways in which users can interact on the platform. Given the time since some of the events in the datasets took place, these changes could have an impact on the results of the network measures. This metric aims to find any correlation between the age of the dataset and the other measures applied in this study, and account for the possible changes in the data. This will hopefully increase confidence in the reliability of the data, especially for the older datasets. With the `created_at` field in the tweet object converted to type `POSIXct`, the mean date for each network is calculated and subtracted from the present. For this study, the 2nd June 2020 is used as the fixed time value.

Attrition In the context of Twitter, attrition refers to the tweets and accounts that have been deleted or privated since the tweet IDs of a dataset have been collected. Due to Twitters terms of service,[2] an account on the platform has the right to delete any tweet they sent or set their account to private, which will either delete that tweet from their database permanently or prevent unauthorised access to that account or tweets via the API. This policy has the effect that while lists of tweet IDs can be shared for academic purposes, its highly likely that some of those tweet IDs no longer retrieve data via the API. This is a particularly important descriptor, since the deletion of tweets or entire accounts could lead to changing the network structure substantially (Zubiaga, 2018). Volkova and Bell (2017, p. 219) suggested that deleted accounts could also signal automated account activity, trolls and other manipulating agents in social media discussions. While these claims are still in the early stages of research, the impact on network structure of deleted or privated accounts can be capture by attrition.

[2]Available at twitter.com/en/tos.

Stellenbosch University https://scholar.sun.ac.za

3.5 Data Analysis

The study has now reached the point where a number of datasets have been gathered, the Twitter discussions have been formatted into a network and various measurements that describe each of the datasets have been obtained. These measurements need to be critically examined to ensure that each variable is independent—that they are not highly correlated with another measure.

The goal is to determine the independent variables in the measurements that will help answer the research question: To indicate whether it is possible to categorise a networked discussion without any domain knowledge. Two particular types of unsupervised machine learning can be applied in succession to achieve this goal. First, *principal components analysis* (PCA), a tool used for dimensionality reduction. Second, *clustering*, a broad class of methods for discovering unknown subgroups in data, as well as determining their quality (James *et al.*, 2013, p. 373). In this section, the necessity of independent variables and methodology for producing those variables will be determined. The clustering method that will be applied to the variables will then be explained together with, the approaches to determining the quantity and quality of clusters obtained from the clustering results. Finally, the detailed requirements for answering the research question will be outlined, for use in the following chapter.

3.5.1 Independence Test

Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of that relationship (Wasserman and Faust, 1994, p. 334). The value of the correlation coefficient varies between 1 and -1. A value on either ends of the extreme, 1 or -1, indicates a perfect degree of association between the two variables. As the correlation coefficient value approaches 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient: a + sign indicates a positive relationship (the variables trend in the same direction) and a $-$ sign indicates an inverse relationship (Wasserman and Faust, 1994, p. 334). There are three standard types of correlation coefficients, namely: Pearson product-moment correlation, Kendall rank correlation and Spearman rank-order correlation. The preferred correlation co-efficient will depend on whether the variables are normally distributed or the relationship between the variables is linear. This will be discussed in more detail in Chapter 4. For this analysis, a correlation matrix is used to visualise the association between each variable. If a strong correlation coefficient is found, then dimensionality reduction will be used to obtain independent variables for the cluster analysis.

Stellenbosch University https://scholar.sun.ac.za

3.5.2 Independent Variables for Cluster Analysis

PCA is a dimension-reduction approach that aims at reducing a large set of variables to a small set that still contains most of the information in the large set (Jolliffe, 2002, p. 1). This tool is widely used in statistical procedures on a wide variety of data for dimension reduction. In the case of this study there are strong correlations found in the results. Specifically, PCA will use an orthogonal transformation to identify principal components, which equal a linear combination of the network measures and are linearly uncorrelated with each other (James *et al.*, 2013, p. 377). The identified principal components are expected to account for most of the variability in the measurements of the network data.

Principal components are the result of trying to find a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. The principal components are calculated by the normalized linear combination of the measurements—also called features—that have the largest variance (James *et al.*, 2013, p. 375). The contribution of each feature to the first principal component is indicated by their specific loading vector. The loadings are constrained, however, so that their sum of squares is equal to one. Otherwise, setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance (James *et al.*, 2013, p. 376). The sum of the products of the normalized features x and their loadings ϕ produce the score given to each principal component.

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip}$$

Once the first principal component Z_1 of the features has been determined, it is possible to find the second principal component Z_2. The second is, again, the linear combination of the features, but now with the maximal variance of all linear combinations that are uncorrelated with Z_1. This is where it becomes apparent that constraining Z_2 to be uncorrelated with Z_1, is equivalent to constraining the direction ϕ_2 to be orthogonal (perpendicular) to the direction ϕ_1 (James *et al.*, 2013, p. 377).

It is important to note that the results obtained from PCA will depend heavily on whether the measurements have been individually scaled (each multiplied by a different constant). This requirement is unique to PCA among other supervised and unsupervised learning techniques, since the features are measured in different units. Some of the variables that will be used are not directly comparable from the respective scores that they produce. If transitivity thus produces a significantly larger numerical score than the other measurements, then the first principal component loading vector will have a very large loading for transitivity, since that variable has by far the highest variance. Because it is undesirable for the principal components obtained to depend on an arbitrary choice of scaling, we typically scale each variable to have standard deviation of one before we perform PCA (James *et al.*, 2013, p. 381).

Stellenbosch University https://scholar.sun.ac.za

3.5.3 Clustering

Clustering will be performed on the principal components to classify the datasets. Clustering refers to a broad set of techniques for finding subgroups in a data set (James *et al.*, 2013, p. 385). The aim of clustering the various networks created from the datasets, is to attempt to partition the datasets into distinct groups, so that the datasets within a group are similar in structure or characteristics, with a high degree of variance between each group. This is an unsupervised problem: trying to discover structure—in this case distinct clusters—based on an unlabelled data set. The difference between PCA and clustering is that PCA attempts to find a low-dimensional representation of the observations that explain the majority of the variance, while clustering hopes to find homogeneous subgroups among the measurements that define each group (James *et al.*, 2013, p. 385).

In this study the clustering method *K-means* is applied to find the clusters that could define each dataset. K-means is a widely used, industry standard approach for partitioning a dataset into a distinct number K of non-overlapping clusters. The mathematical approach that K-means applies results from the idea that a good clustering is one from which the within-cluster variation is as small as possible (James *et al.*, 2013, p. 386). The algorithm attempts to optimize minimal within-cluster variation by determining the mean *squared Euclidean distance* within the features, as defined by James *et al.* (2013, p. 388):

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \epsilon C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \right\}$$

In other words, the within-cluster variation for the kth cluster is the sum of all of the pairwise squared euclidean distances between the observations in the kth cluster, divided by the total number of observations in the kth cluster (James *et al.*, 2013, p. 387).

3.5.4 Determine Cluster Quality

In this study, the optimal number of clusters—and subsequent quality of those clusters—will determine the viability of clustering online discussions around real-world events. The term clustering validation is used to design the procedure of evaluating the results of a clustering algorithm. Two standard metrics, suggested by Tibshirani *et al.* (2001) and Kaufman and Rousseeuw (2013), form the basis of a mathematical as well as graphical interpretation approach. The results from this subsection will form the basis of the requirements to satisfy the research question.

Gap Statistic The *gap statistic* is a standard method, developed by Tibshirani *et al.* (2001, p. 411), for determining the number of clusters in a set of data. This

Stellenbosch University https://scholar.sun.ac.za

method will determine the quantity requirement set for the study in the following subsection. The gap statistic formalises a general approach by using the logarithm of the within-cluster dispersion, by comparing it to its expectation under an appropriate null reference distribution of the data, as represented by Tibshirani *et al.* (2001, p. 412):

$$Gap_n(k) = E_n^* log\big(W_k\big) - log\big(W_k\big)$$

This formula produces the gap statistic score for which the study can solve the quantity requirement. Utilising this method will increase the degree of confidence that the number of clusters generated are as representative of the variance in network structure.

Silhouette Method This study applies the *silhouette method*, a method to determine the quality of clustering and, by extension, the number of clusters that are optimal for the k-means approach. This method will determine the quality requirement set for the study in the following subsection: it determines how well each object lies within its cluster. The graphical method used in this study was discovered by various authors and compiled by Kaufman and Rousseeuw (2013). This method calculates the silhouette width (s) of object (i). Kaufman and Rousseeuw (2013, p. 85) provide the formula: where $a(i)$ is the average dissimilarity of that object (i) to all other objects of hypothetical cluster A and $b(i)$ is the minimum average dissimilarity of i to all objects of hypothetical cluster B.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

When the average of all the silhouette widths has been calculated for the objects in k clusters, this produces a *silhouette coefficient* $\tilde{s}(k)$ per cluster, with the aim to maximise the coefficient:

$$SC = \max_k \tilde{s}(k)$$

The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k. A high average silhouette width indicates a high quality of clustering results (Kaufman and Rousseeuw, 2013, p. 87).

3.5.5 Criteria

To provide a conclusion to this study with a high degree of confidence, clear requirements must be set for the success or failure of the proposed approach. Recall the requirement set by the research question: Can the discussion regarding events on Twitter be *successfully* clustering using network structure?[3] To determine the criteria as inferred by *successfully clustered*, the two aforementioned

[3]Note: the study acknowledges the subjective interpretability of *successful* in this context. There is no objective defined methodology to determine whether or not the analysis is successful.

Stellenbosch University https://scholar.sun.ac.za

methods will be used. They provide a computational solution to determine the optimal quantity of clusters and the quality of the structure of those clusters. The values of where their results coincide will produce the optimal amount of high structure quality clusters. The results produced by the clustering methods provide the grounds on which the success of the clustering approach will be determined. The purpose of this study has always been to provide a proof of concept for utilising network structure to cluster social media data, with the important limitation of no prior domain knowledge. It would thus be unrealistic to claim that a measure of accuracy can be determined without labelled data with which to compare the results. The criteria set out in this section will serve to satisfy the requirements of determining the viability of utilising network measures to cluster social media discussions. The future work to refine this approach and expand its effectiveness will be discussed in Chapter 5.

Given that there are 39 datasets, (see Chapter 4 section 4.2) this provides the study with a ceiling value for the amount of clusters. The study would thus be considered a failure if the gap statistic score and silhouette coefficient are higher than 39. Similarly, if the result is 1, then the methods failed to produce any interpretable clusters. With regards to the quality of the clusters, Kaufman and Rousseeuw (2013, p. 88) state that the interpretation of the silhouette method results may vary when applied to real-world data. But in their experience they provided a proposed interpretation of the results: A silhouette score between 0.26 and 0.50 suggests that the structure of the cluster is weak and could be artificial, with a score lower than 0.25 inferring that no structure is found. Any result for the number of clusters proposed by the gap statistic method with these silhouette scores would support an outcome in which no successful clustering is possible.

The results from the tests will need to coincide in order to determine a reasonable structure quality for the study. Consider the possible outcomes of the cluster tests for the interpretation of the results. The study would be rejected if either of the method's results fall outside the requirements. However, it is possible that both methods produce results that fall within the requirements, but that the range of values to not overlap. For example, the optimal amount of clusters, as determined by the gap statistic method, could be 25 clusters, but the silhouette score is maximised for five to ten clusters. While this would not result in an outright reject for the viability of the approach, it is not an ideal outcome. The ideal outcome would involve the results of both methods falling in the same range, where one or two values can be defined as the optimal clusters with reasonable structure quality.

Therefore, a range of values for a high degree of confidence is also required to satisfy the research question. The quantity requirement for success is based on

This study rather strives to match the outcome of this study to the criteria specifically outlined in this section. These criteria are defined to verify the proof of concept: to classify social media data, by means of clustering, by utilising network structure derived from interactions on Twitter.

Stellenbosch University https://scholar.sun.ac.za

the search parameters that were used to collect the data. Table 1 in the Appendix provides a complete set of descriptives for each dataset in the study, as well as the context for the search parameters used to collect the data. The amount of optimal clusters will be used to determine if the study is rejected. But similarly, where the amount coincides with clusters of reasonable quality, a higher degree of confidence will be placed in the approach. A more detailed discussion regarding clustering results will be had in the following chapter. With regards to the quality of the clusters, the study refers to the interpretations suggested by Kaufman and Rousseeuw (2013, p. 88): A silhouette score between 0.51 and 0.70 is proposed that a reasonable structure quality has been found in the results, with a value higher than 0.71 indicating a strong structure quality. These silhouette scores for the optimal amount of clusters, as determined by the gap statistic results outlined by Kaufman and Rousseeuw (2013, p. 88) will lead the study to accept that the data can be clustered using network measures and descriptors.

3.6 Ethical Considerations

For the purpose of this study, the public domain tweets have been stripped of any further elements that may be considered personal information or opinions, since they are not used in this approach. The tweet text, any geo-tagging, user details or any other identifying information have been removed. Only the interaction data was used to create a network of the discussion. Furthermore, once the measures have been applied and the results are obtained, the networks themselves were also discarded. The analysis and discussion were only performed on the measurements taken from the networks. Ethical approval in the form of exception from full scrutiny was obtained, since none of the tweet IDs were collected firsthand, all the data utilised by this study is non-personal and in the public domain.

3.7 Chapter Summary

In this chapter, the methodology for the proposed study was explored in detail. From the data options discussed in Chapter 2, the source was chosen and the method for collecting the data was defined. The interactions contained within the Twitter datasets was formatted to convert accurately the types of interactions into directed edges between actors in a network. These interactions were modelled and prepared for the network measures that will be applied in the following chapter. The data analysis was discussed as well as the specific methods whereby the measures will be calculated and the clustering analysis be performed. Finally, a set of tests and criteria were defined with which the research question will be addressed, as well as a brief description of the ethical considerations considered for this study.

Stellenbosch University https://scholar.sun.ac.za

Chapter 4

Analysis and Findings

4.1 Chapter Introduction

Chapter 2 introduced the idea that humans are embedded in social networks. They are dependent on the structure of these networks at both group and individual levels to comprehend and explore the informal communities that support society. Similarly, social media and information that is distributed by other media to the broader populace could have a lasting effect on the actions taken by users offline. This is the motivation for understanding the digital conversations that take place on a variety of online platforms. This study chose Twitter as the source for these interactions, due to its popular use by media as a reference and it's open form of discussion (Paulussen and Harder, 2014, p. 8). Chapter 3 laid the groundwork for the necessary steps to analyse the data obtained from this source. The primary research question concerns the possibility of clustering the event-discussions by finding clusters in data, utilising network measures that create a defining structure of any given event online. Since the network measures objectively characterise the event, little or no domain knowledge is thus required to differentiate events. To determine if this is the case, there are two groups of variables to extract from the data: *network* measures and *descriptors*. Measurements for networks are a set of graph-level indices that give insight into how the network is structured; where the co-variate network measures account for the difference in size; and the time and age of the data that affect the network measures. The correlation of these groups of variables attempts to produce more reliable results, despite the diverse nature of the datasets. Once the set of dimensions have been determined, they will be clustered, and a set of measures will be applied to determine the optimum result.

Once the analysis has been completed, the research question will be answered in the findings section of this chapter. With the goal of the study satisfied, a more detailed analysis of the results will be explored further in the findings. The first step, however, is to describe the data as it has been gathered from the source, and how these statistics might affect the results.

Stellenbosch University https://scholar.sun.ac.za

4.2 Descriptive Statistics

The descriptive statistics for the discussions that took place on the platform are embedded in the data and may give insight into the results of the analysis. It is therefore important to find and highlight any discrepancies that may affect how the discussion is structured. As explained in chapter 3, a total of 39 datasets were collected from the two primary sources—25 from Cornelissen *et al.* (2019) and 14 from Zubiaga (2018). Note that Zubiaga (2018) has 30 data sets in their final study, but only 14 were used due to the technical constraints of working with this amount of data. Table 4.1 provides an overview of the key descriptive statistics for all 39 datasets.

Table 4.1: Descriptive Statistics for the 39 Datasets

Descriptive	*Mean*	*Std dev*	*Median*	*Max*	*Min*
Original Tweet Count	762267	1039599	168529	5044378	17238
Tweet Count	232523	268481	89303	1106427	7504
User Count	179958	251326	54076	1246776	8076
Attrition (%)	62.67	12.30	66.00	78.00	11.00
Capture Range (days)	28.26	27.66	17.80	97.57	1.22

As shown in Table 4.1, the datasets ranged from a maximum of 5044378 tweets for the largest dataset, to 17238 tweets for the smallest. However, some tweets for each dataset have since been deleted or the accounts that sent the tweets were made private.[1] The largest dataset used in the study contained 1106427 tweets, and the smallest consisted of 7504 tweets after the rehydration process.[2] An average of 179958 users were found in each dataset used in the study. The datasets have an average rate of attrition of 62.67%, and were captured over a median of 17.80 days per dataset. As explained in Chapter 3, the process that Twitter requires for data gathering results in a loss of all the tweets that have since been deleted. The attrition row notes the percentage of tweets lost for each dataset. It makes sense that the older the dataset is, the more tweets are likely to be deleted, as indicated in the start date column in Table 1. This result indicates, however, that the relationship between age and attrition is not linear, as seen in the correlation results on Table 4.3. All the datasets for this study have been collected before September 2018. At the time of writing they are therefore a minimum of 21 months old. The attrition is accounted for as a descriptor and added as a measure to the analysis.

Table 4.2 provides an overview of the key descriptive statistics for each measure calculated from the dataset. As would be expected with network data, most

[1] When a Twitter account has the private setting enabled in their privacy options, it restricts the visibility of that account's tweets to the followers of that account.

[2] Further detail for each dataset is available in Table 1 in the Appendix, sorted by the number of tweets each dataset had originally before they were rehydrated with the Twitter API.

Stellenbosch University https://scholar.sun.ac.za

of the data is highly positively skewed, except for diameter, eccentricity and data age.

Table 4.2: Descriptive Statistics for the 14 Network Measures and Descriptors.

Measure	*Mean*	*Std dev*	*Median*	*Min*	*Max*	*Skew*	*Kurtosis*
Transitivity	0.60	0.79	0.38	0.01	3.89	2.44	6.24
Reciprocity	0.71	0.69	0.54	0.01	3.83	2.53	8.17
Centralisation	0.60	0.79	0.23	0.04	3.15	1.78	2.19
Closeness	0.48	0.87	0.13	0.01	3.52	2.42	4.89
Density	0.51	0.86	0.16	0.02	3.04	2.03	2.77
Assortativity	0.21	0.98	−0.01	−1.16	4.54	2.51	8.16
Betweenness	0.40	0.91	0.03	0.00	3.91	2.71	6.72
Diameter	0.93	0.34	0.93	0.15	1.86	0.38	0.54
Eccentricity	0.94	0.30	0.86	0.46	1.65	0.31	-0.83
Capture Range	0.71	0.69	0.45	0.03	2.45	1.17	0.30
Node Count	0.58	0.81	0.17	0.03	4.02	2.29	6.17
Edge Count	0.65	0.75	0.26	0.02	3.09	1.34	1.40
Data Age	0.91	0.39	0.93	0.43	1.88	0.86	0.04
Attrition	0.96	0.25	1.03	0.11	1.23	−1.77	3.24

Note The measures are scaled once all the results have been compiled. All measures produce a value greater than 0, except for assortativity that ranges between −1 and 1 unscaled. An unscaled edition, Table 2, is found in the Appendix.

For this study, the datasets have been anonymised. One of four categories assigned to each dataset, based on the search parameters, have been assigned depending on the context in which the data was gathered.[3] The categories are: *Politics SA* for datasets that were gathered regarding political discussions and events that took place in South Africa, from Cornelissen *et al.* (2019); *Politics Int*, representing datasets that were gathered in a political discussion for events that took place outside of South Africa; *Disaster* for datasets gathered with natural disaster search parameters, for example “fire”, “hurricane”, “earthquake”, “flood” and more; and finally *Social* representing datasets that revolved around events that were discussed on Twitter about sports or entertainment etc. from Zubiaga (2018).

4.2.1 Bi-variate Correlations

Broad descriptive and summary statistics for the data were provided in the previous section. The interaction between these variables should be explored fur-

[3]For this and following sections, events that were collected with similar areas of interest, will be referred to, as being in the same *category* or having the similar search parameter contexts. This is not an attempt at manual labelling, but rather that the search parameters were similar in the type of event searched.

Stellenbosch University https://scholar.sun.ac.za

ther to investigate observable patterns and possible co-linearity between variables. Given that the majority of the data is skewed (as indicated in Table 4.2), noting the recommendation made by Hauke and Kossowski (2011, 89) was followed, namely that the Spearman rank-order correlation method is more capable of handling data that is not normally distributed. This correlation coefficient was thus used for the bi-variate correlation analysis.

Table 4.3: Correlation Table for Measures

Measure	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1. Trans													
2. Reci	.37†												
3. Centr	−.20	.18											
4. Close	−.30	.17	.71										
5. Dens	−.30	.30	.71	.94									
6. Assort	.34†	.20	−.11	.15	.17								
7. Betw	−.40†	.19	.74	.89	.90	.03							
8. Diam	−.37†	−.12	.21	.62	.54*	.24	.61						
9. Ecce	−.43$^{\triangleleft}$	−.15	.14	.52*	.45$^{\triangleleft}$	.07	.56*	.70					
1. Range	.10	.36†	.23	.09	.09	−.18	.04	−.21	−.11				
11. Node	−.53*	−.10	.45$^{\triangleleft}$	.75	.79	.15	.79	.68	.75	−.13			
12. Edge	−.34†	.18	.63	.88	.94	.24	.90	.64	.61	−.06	.89		
13. Age	−.32†	−.04	.10	.43$^{\triangleleft}$	.52*	.39†	.48$^{\triangleleft}$	.63	.57*	−.35†	.72	.66	
14. Attr	−.35†	−.62	.02	.18	.13	.04	.15	.49$^{\triangleleft}$	.52*	−.21	.55*	.25	.47$^{\triangleleft}$

Note $* = p < .001$; $\triangleleft = p < .01$; $\dagger = p < .05$

The correlation test reveals moderate to strong correlations between several network measures, as well as the co-variate network measures, with ten being statistically significant at a p value of 0.05 for the † level, seven for a p value of $0.01(\triangleleft)$ and eight for a p value of $0.001(*)$. Given these results, the need for a new set of independent variables is evident. The study therefore continued with applying principal component analysis to obtain independent variables.

4.3 Analysis

Recall in Section 3.5 of Chapter 3, the study has reached the point where all the data preparation has been planned, and the steps for the analysis could be determined. Similarly, in this chapter, all the following results will be used in Chapter 5 for the discussion.

4.3.1 Producing Principal Components

The starting point of PCA is the matrix of correlation coefficients derived from the original data set. The rationale behind this method requires that the correlations be obtained from variables measured on a continuous scale. The above sections

Stellenbosch University https://scholar.sun.ac.za

produced these correlations and provided the PCA method with the necessary values. Given the input of the original 14 variables listed in Chapter 2, the results of this method reduces the multidimensional datasets to a lower number of dimensions for further analysis and incorporates the correlation between measures in creating new dimensions. First, the study must determine the number of principal components (PC) to retain. To assist with this choice, the method produces a scree plot. The procedure of finding statistically significant factors or components using a scree plot is also known as a scree test (Cattell, 1966, p. 246). The scree plot, depicted in Figure 4.1, plots the percentage of explained variance for each of the new principal components.[4]

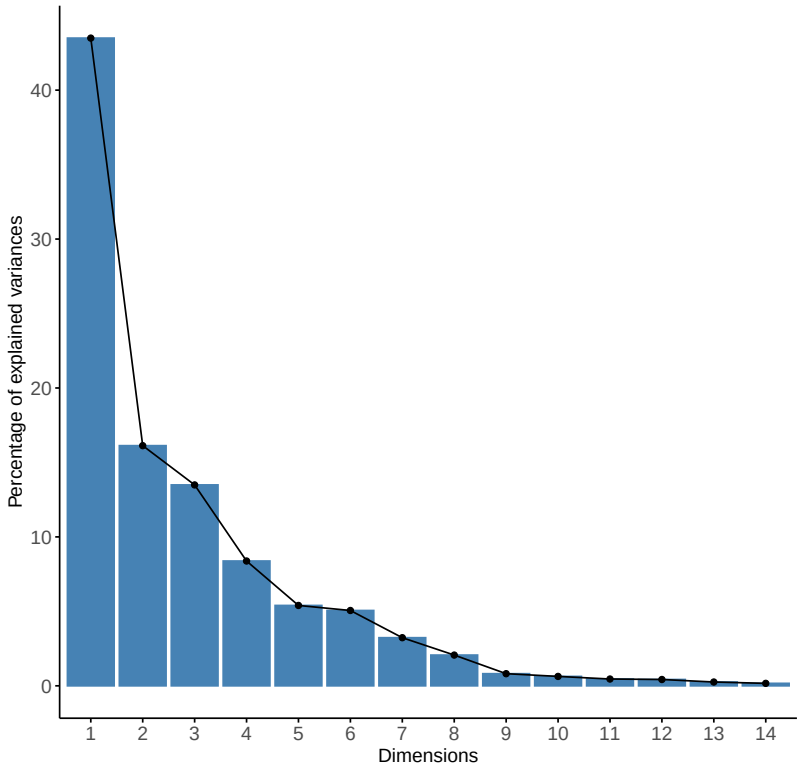


Figure 4.1: Scree plot

Note: Scree plot indicating proportion of variance for each component

As is evident in Figure 4.1, PC 1 accounts for 43.50% of the variance. Following this, the results of the scree test indicate substantial drops in explained variance from the first to the second dimension, the third to fourth and, finally, the sixth to

[4]Scree plot is a line plot of the eigenvalues of factors or principal components in an analysis, introduced by Cattell (1966). The y-axis captures the eigenvalues that indicate the proportion of variance contributed by each principal component.

Stellenbosch University https://scholar.sun.ac.za

seventh dimension. The results of the scree test would therefore result in choosing either two, four or six dimensions to use in the rest of the analysis. For this study, a cut-off of 90% of the variance in the data has been chosen as a suitable variance to retain. Consequently, six dimensions, which account for 91.95% of the total variance, were retained for the subsequent analysis procedures.

The next step is to inspect which measures contribute to these six dimensions. The PCA method produces a set of components, or factor loadings that indicate how much each measure contributed to the dimensions in percentages.

Table 4.4: Factor Loadings for Network Measures and Descriptors

Measure	*PC 1*	*PC 2*	*PC 3*	*PC 4*	*PC 5*	*PC 6*
Transitivity	0.48	3.00	**34.24**	**16.68**	4.31	**34.32**
Reciprocity	0.08	5.04	8.60	**25.55**	**18.24**	**16.76**
Centralisation	9.27	**10.75**	1.22	**17.11**	1.07	1.58
Closeness	**20.75**	1.90	0.87	1.85	9.23	1.37
Density	**19.49**	1.98	4.10	0.50	7.84	0.39
Assortativity	0.00	**43.34**	**39.37**	9.36	0.85	3.94
Betweenness	**20.04**	0.03	0.16	**10.74**	0.11	**15.99**
Diameter	0.99	2.67	0.00	0.36	1.75	0.62
Eccentricity	0.74	1.74	0.44	0.26	2.62	0.22
Capture Range	0.02	**15.32**	4.57	**11.50**	**31.50**	0.28
Node Count	**11.67**	7.00	6.27	1.29	**16.00**	**17.60**
Edge Count	**14.73**	2.57	0.02	2.56	3.38	5.06
Data Age	1.59	3.56	0.02	2.19	0.35	0.03
Attrition	0.16	1.09	0.10	0.06	2.76	1.84
Eigenvalue	3.03	1.12	0.94	0.58	0.38	0.35
Variance %	43.50	16.13	13.49	8.38	5.40	5.06
Cum. Variance %	43.50	59.62	73.11	81.50	86.89	91.95

Note Loadings higher than 10 are in bold, indicating substantial contributions in variance.

The variance percentage in Table 4.4 indicates how much each component explains the total variance found in the data. This is visualised in Figure 4.1 above. The value determines what amount of components to retain. Table 4.4 showcases the loadings, where the highlighted percentages indicate a substantial contribution—higher than 10%—to the component. As is evident for PC 1 the centrality measures, betweenness, closeness and, to a slightly lesser degree, centralisation account for over 40% of the component and are accompanied by density, node and edge counts. From their explanations in Chapter 2, it is known that these measures are closely related and correlate with the size of the network, calculated with the node and edge counts. Therefore PC 1 is affirmed by the theory and high correlation between these measures. Interestingly, for PC 2, assortativity is by far the majority contributor, with centralisation adding the rest of network measure contributions to the component. However, the range in number of days in which the data was captured also seems to have affected these measures, such that it contributes 15.32% to this component. This indicates that

Stellenbosch University https://scholar.sun.ac.za

the assortativity measure provides a unique insight into the structure of a network and explains a substantial 16.13% of the variance in the data. But high correlation between these measures and the capture range means that it had to be accounted for. For PC 3, transitivity and assortativity accounts for 73.61% of the component. This suggests that these measures have a significant ($p < 0.05$) amount of correlation and, according to the definitions in Chapter 2, this would ring true. Both measures attempt to capture characteristics of homophily, where actors interact with one another, based on a similarity co-efficient. PC 1 through 3 account for 73.11% of the variance found in the data. This is substantial, but not enough for the 90% cut-off requirement. PC 3, 5 and 6 indicate that all the measures already discussed in this paragraph, further contribute to these measures, respective to how the measures correlate. It is noteworthy that diameter, eccentricity, data age and attrition did not substantially contributing to any of the six listed components. This is not an unexpected outcome. In Chapter 3, the necessity of determining the usefulness of some measures was anticipated. This outcome will be discussed in more detail in Chapter 5. Adding the descriptors to the list of variables for the study has shown that some contribute notably to the variance in these components. By including this, any variance in the data due to the descriptors of the data will be accounted for. Figure 4.2 illustrates the variable contributions of each measure to the principal components. It depicts the first two principal components and the direction and contribution of each measure.

From Figure 4.2, the loadings explored in Section 4.3.1 are evident. The centrality measures, density, node and edge counts contribute roughly equal amounts to PC or Dimension (Dim) 1,[5] and assortativity is the majority contributor to PC 2.

4.3.2 Cluster Quality

The six dimensions obtained from the PCA are used to explain the variance between all the datasets. A clustering method is applied to determine clusters with similar variance. Two methods are used to determine the amount and quality of the clusters.

First, the gap statistic method is applied to account for the optimal amount of clusters. As explained in Chapter 3, the gap statistic method is used to calculate the optimal amount of clusters to generate using the *K-means* clustering method. Since the amount of clusters generated by K-means is user-defined, the aim is to find the amount that most clearly separates the networks, based on the variance in their structure. To obtain an ideal clustering, Tibshirani *et al.* (2001) explains that one should select k such that you maximize the gap statistic.

As depicted in Figure 4.3, the cadence of the plot is always increasing. It is common to find that in many real-world datasets, the clusters are not as well-

[5]The terms 'principal components' and 'dimensions' have the same meaning and are used interchangeably in this paper.

Stellenbosch University https://scholar.sun.ac.za

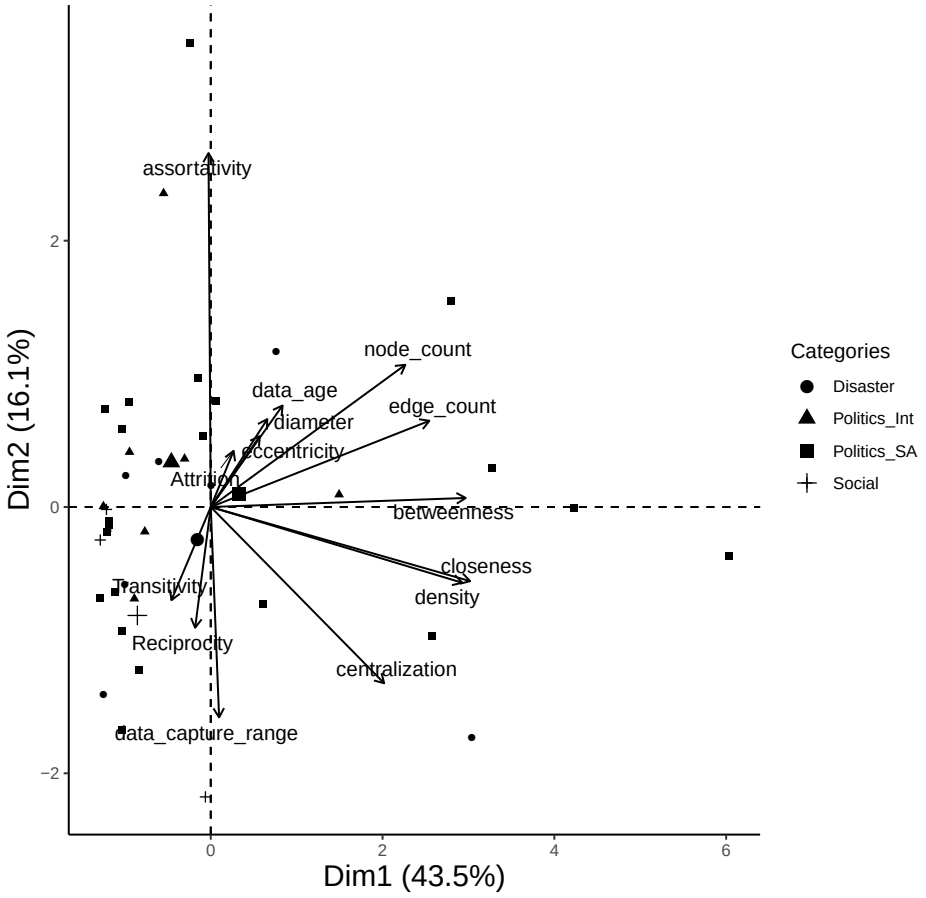


Figure 4.2: Bi-plot for PC 1 and 2

Note: The centroid of each category is indicated by larger symbol of that category.

defined. To account for this, Tibshirani *et al.* (2001) suggests the 1-standard-error method:

Choose the cluster size k to be the smallest k such that:

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

In other words, identify the point at which the rate of increase of the gap statistic begins to reduce. From Figure 4.3, the amount of clusters that satisfy this criteria, is seven or eight before the score decreases.

Second, the silhouette method is applied to account for the quality of the clusters. The silhouette method, as explained in Chapter 3, is used to calculate the quality of clusters generated with the *K-means* clustering method. This method also further affirms the amount of clusters that should be generated, as the cluster amount with the highest quality is preferred.

As depicted in Figure 4.4, the highest average silhouette width is found at five clusters, and the width substantially decreases after eight clusters. According to

Stellenbosch University https://scholar.sun.ac.za

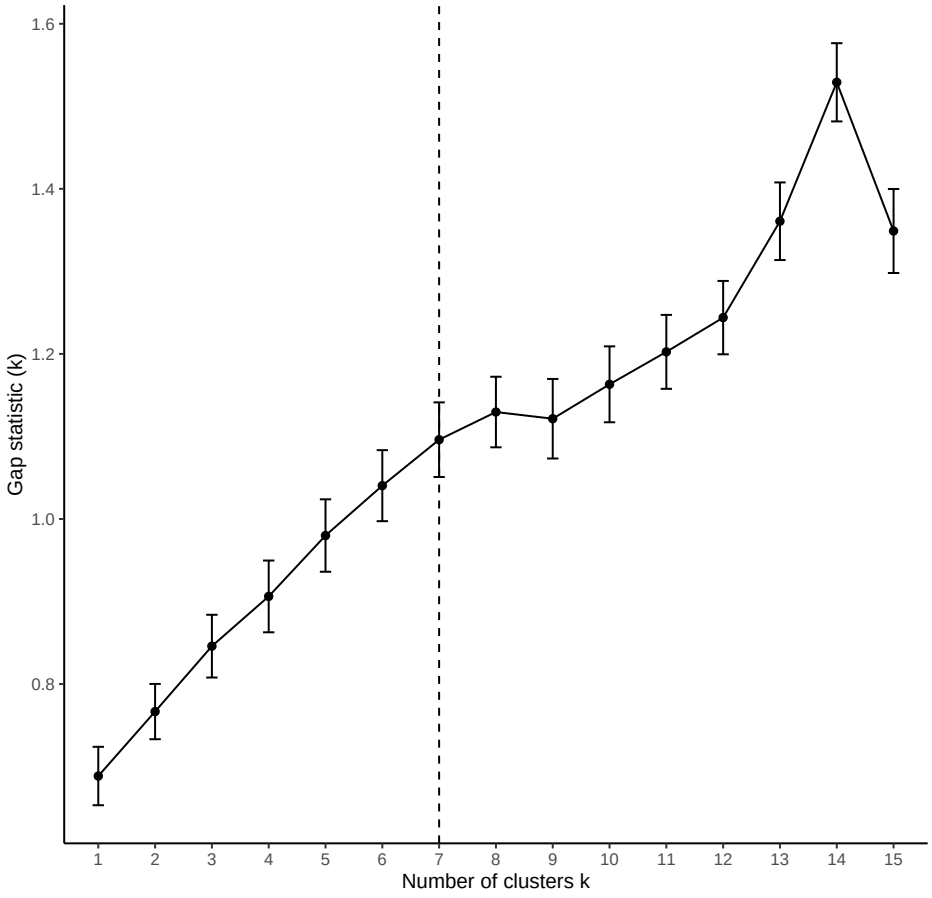


Figure 4.3: Optimal Number of Clusters as Determined by Gap Statistic Method

Kaufman and Rousseeuw (2013, p. 88), a Silhouette coefficient between 0.51 and 0.70 proposes that a reasonable structure quality has been found. All the values between two and eight clusters fall within this bracket, with a high of 0.60 at three and five clusters, and low of 0.57 at eight clusters.

4.4 Findings

To evaluate the findings, consider the requirements set by the research question. Can the discussion regarding events on Twitter be successfully clustered by measuring and classifying them by utilising the network structure derived from interactions revolving around those events, without prior domain knowledge? The requirement states that the events be clustered successfully, which constitutes an amount of clusters less than the original amount of events, and that those clusters have a reasonable structure quality. The greater the difference in amount of clusters found with regard to the original dataset count, within the boundaries set in Chapter 3, and a higher silhouette coefficient for those clusters will deter-

Stellenbosch University https://scholar.sun.ac.za

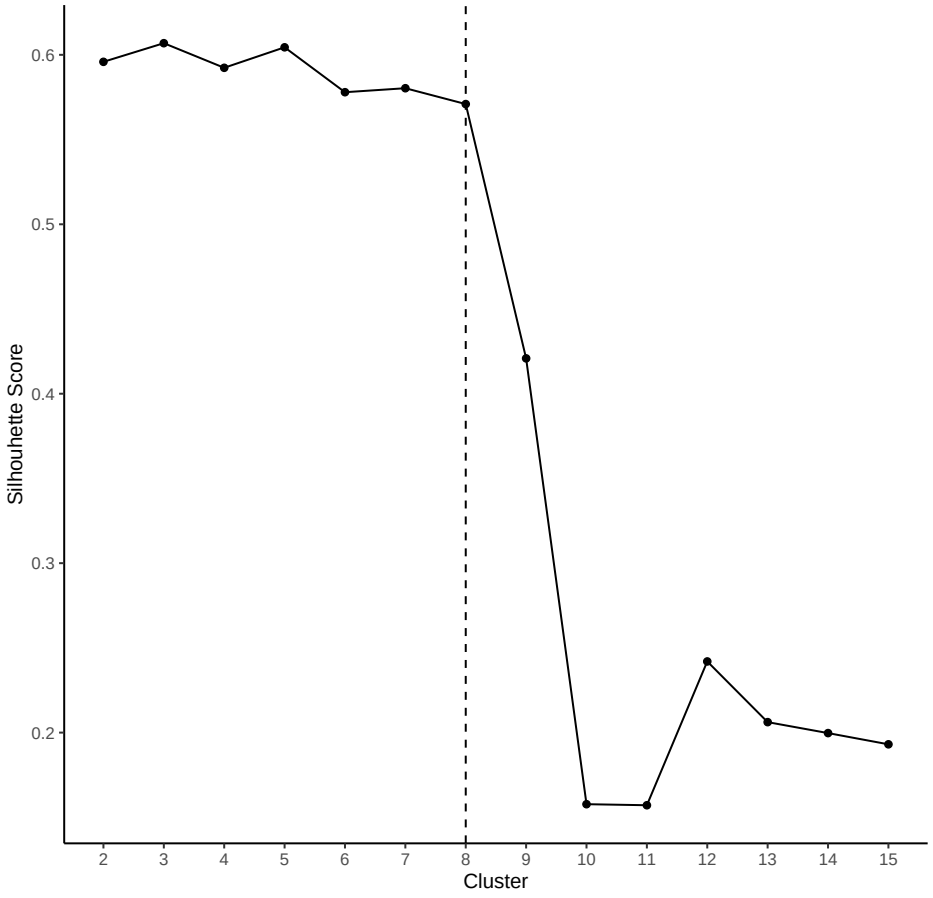


Figure 4.4: Quality of Structure per Cluster

mine the degree of success for the study.

The result will be determined from where the values produced by the two methods in the previous section coincide. The gap statistic method determined that seven clusters would optimal. The silhouette method produced a reasonable structure quality score peak for five clusters, but also had reasonable scores up to eight before values decreased substantially. This leads the study to conclude that up to eight high quality clusters can be found in 39 total datasets, with an optimal count of seven or eight clusters. This results in a final count of seven or eight clusters for the consideration of the research question.

Therefore, given the criteria as defined in section 3.5.5, the study concludes that it is possible to successfully cluster data based on network structure and without prior domain knowledge.

Stellenbosch University https://scholar.sun.ac.za

4.4.1 Clustering Results

With the positive result of the clustering of the data, the study proceeds to explore the variations that are found in interesting subsets of the data. Table 4.5 lists the 39 datasets, ordered by which cluster they occupy, with their respective PCA scores. For the purpose of this section, a k value of eight was chosen to generate clusters. This decision is based on the assumption that more clusters will offer increased granularity when analysing the clustering results. Each cluster has a centroid within the six dimensions and the euclidean distance from each dataset to the centroid has been calculated. This measure will indicate how representative a dataset is of the cluster's characteristics: a lower distance is more representative. These details will be discussed more fully later in this section, by visualising the clusters pairs of principal components and referring to Table 4.4 as to which network measures contributed to the relevant components.

From Table 4.5, it is evident that events with similar categories do not strictly fall in the same clusters. This indicates that the approach identified other patterns not easily identify-able without domain knowledge and further analysis. Some interesting observations from Table 4.5 are the exemplar datasets: those most representative of each cluster. While it is not possible to visually analyse a six-dimensional plot of the eight clusters, it is feasible to analyse the bi-variate visualisations of the 39 datasets on pairs of the principal components.

Consider the first cluster in Table 4.5. The datasets found in cluster 1 already diverge in terms of categories but, given that there are two where their PC scores are similar would indicate the average characteristics of cluster 1. Note that this observation would also apply for cluster 3 and 5, where there is only one dataset in cluster 3, and two in cluster 5. In this case, however, the datasets in cluster 1 have similar results for PC 3 and, to a lesser degree, PC 4 and 6. To explore these phenomena, Figures 4.5 and 4.6 highlight pairs of principal components and where the clusters are found on the axes, also referred to as biplots. Note that for Figures 4.5b, 4.6a and 4.6b, cluster 1 is a clear outlier. A complete set of the dataset clusters, visualised on pairs of principal components is found in the Appendix at Figures 1–8.

Figure 4.6b depicts PC 4 and 6 on the same Cartesian plane, with the clusters colour coded, the categories in unique shapes and the measure contributions for those components directed and weighted. The fourth and sixth dimensions have both the most similar and highest scores for cluster 1. Table 4.4 indicates that reciprocity, centralisation, transitivity, betweenness and capture range contributed substantially to PC 4. Transitivity, reciprocity and betweenness also contributed to PC 6, with the addition of the node count. This is supported by Figures 4.5 and 4.6 in all three pairs of principal components. However, since a measure can contribute negatively to a component, it is evident that betweenness and reciprocity were the most positive contributors. This would seem to suggest that the variance of the network structure of cluster 1 is largely explained by these measures.

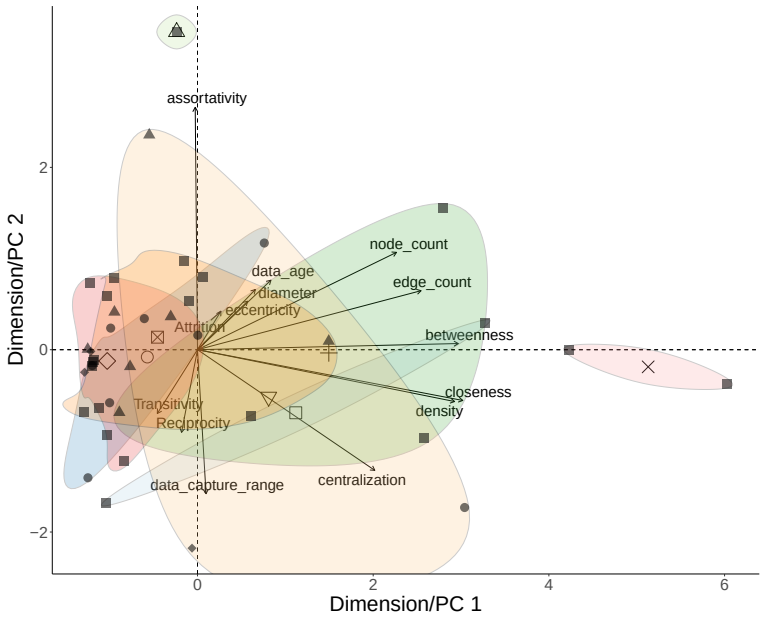
Stellenbosch University https://scholar.sun.ac.za

Table 4.5: Cluster Results for 39 Datasets

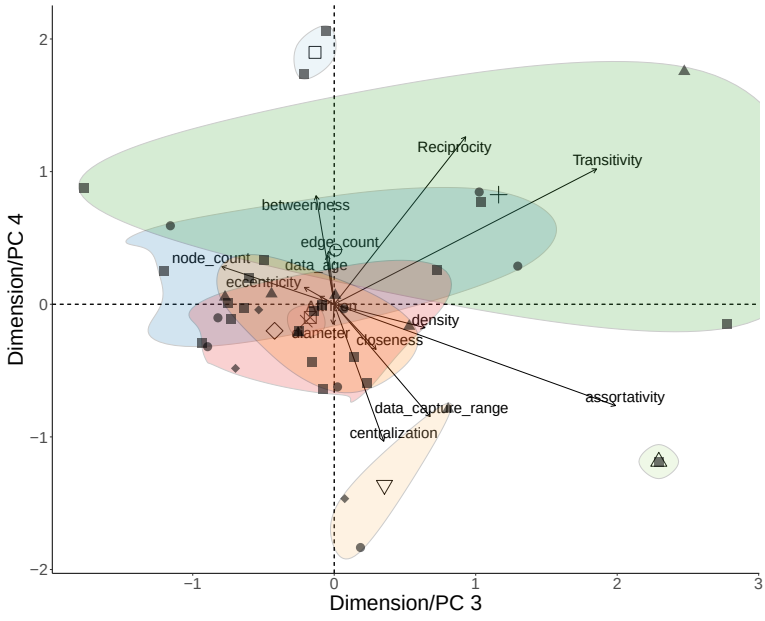
Cluster	Category	*PC 1*	*PC 2*	*PC 3*	*PC 4*	*PC 5*	*PC 6*	*Eucl. Dist.*
1	Politics_SA	0.91	6.42	0.12	13.30	8.96	16.43	**2.51**
1	Social	9.08	0.20	0.01	18.61	0.08	23.17	**2.51**
2	Politics_SA	0.31	0.27	4.59	0.36	0.42	0.06	3.51
2	Disaster	0.01	0.64	3.94	0.27	0.03	1.31	3.32
2	Politics_SA*	1.23	0.08	2.92	2.60	0.11	0.04	**2.54**
2	Social	0.49	3.13	3.66	1.54	4.32	3.72	7.32
2	Politics_SA	0.90	1.98	2.38	0.37	0.41	0.52	2.94
2	Politics_SA	1.32	4.51	2.87	3.15	1.37	1.30	4.99
3	Disaster	0.05	27.75	14.36	6.18	7.60	4.41	**0.00**
4	Disaster	6.61	5.48	8.51	3.35	17.02	13.83	8.94
4	Politics_Int	5.63	2.14	21.05	0.10	1.33	4.72	10.89
4	Politics_Int*	0.67	1.08	16.71	13.53	0.11	15.71	**1.95**
5	Politics_Int	30.78	0.31	0.06	0.01	1.23	0.67	**3.03**
5	Politics_Int	15.15	0.00	0.17	0.18	10.46	1.04	**3.03**
6	Politics_SA	1.40	0.14	0.78	0.01	3.19	0.00	1.05
6	Politics_SA	1.19	0.04	1.46	0.06	1.90	0.02	1.90
6	Politics_SA	0.83	0.13	1.84	0.05	1.73	0.01	1.98
6	Politics_SA	1.18	0.03	1.10	0.00	2.30	0.04	1.91
6	Disaster*	0.00	0.06	2.19	0.45	2.44	1.48	**0.05**
6	Politics_SA	1.32	0.00	1.63	0.01	5.14	0.04	1.57
6	Politics_SA	1.06	0.93	0.15	1.53	2.21	0.02	0.67
6	Social	1.25	0.00	1.33	1.03	1.46	0.27	1.23
6	Politics_SA	1.27	1.23	1.43	0.29	1.79	0.13	0.44
6	Politics_SA	0.59	3.42	0.02	1.79	6.70	1.62	7.56
7	Disaster	0.26	12.67	1.77	2.74	0.01	0.88	8.76
7	Politics_SA	0.00	10.82	0.02	9.44	10.98	0.47	4.64
7	Politics_Int*	7.83	6.84	0.09	14.78	0.96	0.72	**4.12**
8	Politics_SA	0.31	1.22	0.07	0.85	1.30	0.73	1.08
8	Politics_SA	1.22	0.07	0.01	0.01	0.00	0.32	1.77
8	Politics_SA	0.90	0.79	0.05	0.68	0.51	1.24	0.79
8	Politics_SA	0.76	0.39	0.00	0.02	0.05	0.43	1.74
8	Disaster	0.76	1.41	0.02	0.00	0.94	0.03	0.23
8	Politics_SA	0.50	0.08	0.18	0.21	0.60	1.01	0.82
8	Politics_Int*	1.89	0.02	0.77	0.12	0.47	0.11	**0.01**
8	Politics_Int	0.00	1.44	1.00	0.17	0.01	1.96	1.19
8	Disaster	0.02	2.15	1.53	0.00	0.00	0.75	1.07
8	Politics_SA	0.85	0.77	0.00	1.71	0.49	0.17	0.60
8	Disaster	0.08	0.30	0.54	0.03	0.18	0.42	1.84
8	Politics_SA	1.40	1.08	0.67	0.49	1.19	0.23	1.68

Note Event with the lowest euclidean distance to the centroid are bold, indicating that the marked datasets (*) are most representative of that cluster, given that there are three or more clusters. If there are only two clusters in a dataset, they would both be equally far away from the centroid

Stellenbosch University https://scholar.sun.ac.za



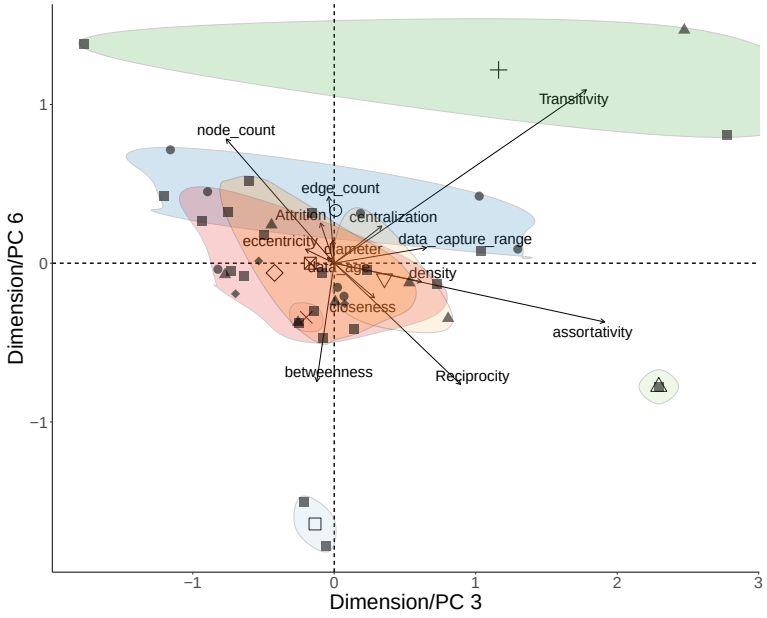
(a) PC 1 & 2



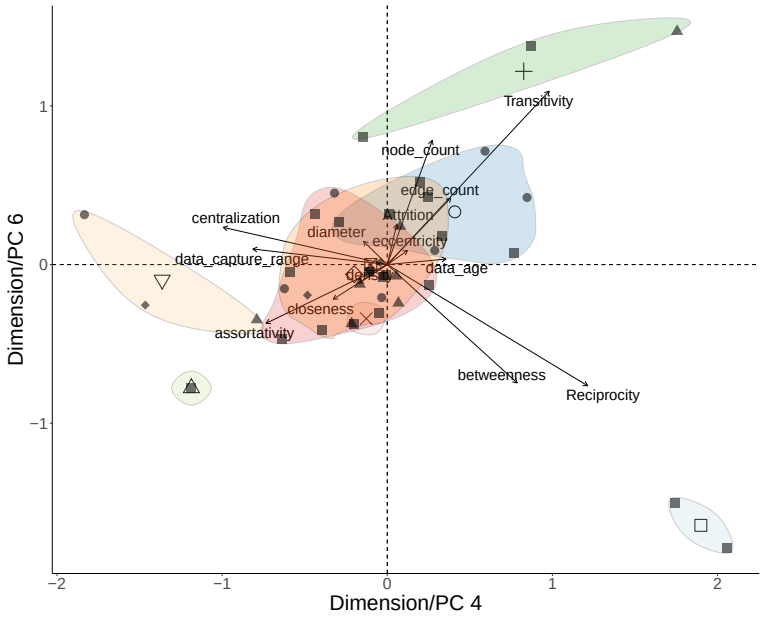
(b) PC 3 & 4

Figure 4.5: Cluster 1: Bi-plot and Clusters (PC 1 & 2, PC 3 & 4)

Stellenbosch University https://scholar.sun.ac.za



(a) PC 3 & 6



(b) PC 4 & 6

Figure 4.6: Cluster 1: Bi-plot and Clusters (PC 3 & 6, PC 4 & 6)

Cluster 1: All the combinations of principal components that have similar or high PC scores for cluster 1. All the scores in the bi-plots for this section are scaled to fit the datasets more practically on the same axes. The cluster shape position indicates the centroid of each cluster.

Stellenbosch University https://scholar.sun.ac.za

This is supported by Figure 4.5a. Since the overall variance in the data is largely explained by PC 1, the most substantial differences would be expected in this cluster. The datasets in cluster 1, however, are still definitively explained by betweenness and reciprocity. By using these two measures, the approach was able to separate completely cluster 1 from other clusters.

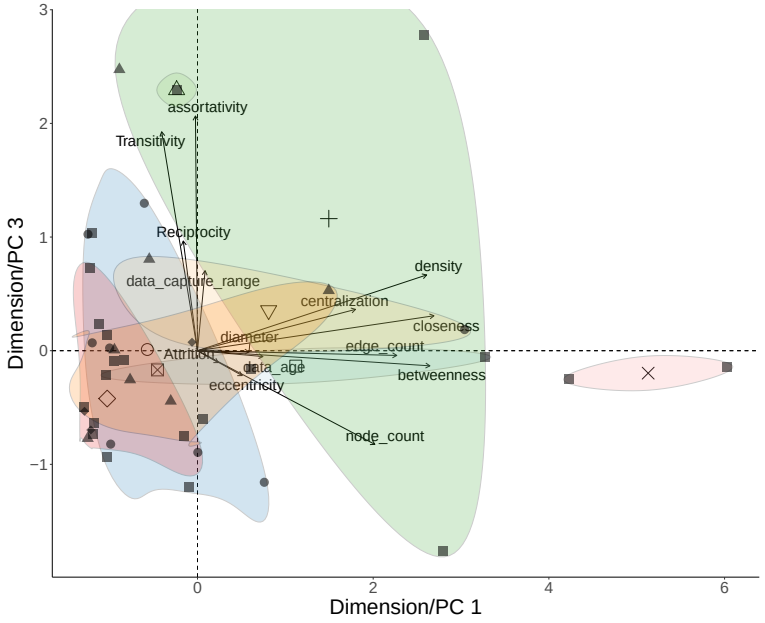
Cluster 2 provides more variety in terms the amount of datasets found in the cluster. While the majority of categories are from South African political events, a disaster and social event is also included. This is where the euclidean distance from the datasets to the centroid will give addition insight as to how the cluster is structured. First, a South African political dataset is closest to the centroid, hinting that the cluster is structured around this category of events, with a maximum distance deviation of 2.45 from the exemplar dataset. The outlier here is indicated. It has highest distance from the centroid and is found in the social category of events. Consider the PC scores in Table 4.5. The scores are representative of the measures that define the network structure and of all datasets in cluster 2. They have a similar score for PC 3. The third principal component is composed largely of transitivity and assortativity, as seen in Table 4.4. While there are no other obvious similarities in cluster 2 for the PCA scores, some of the datasets that have a similar euclidean distance do display similar PCA scores. The first two datasets in cluster 2 have a distance variance of 0.19, with PC 2 and 4 scores that indicate equally low degrees of variance. In Figure 4.7, most of the datasets in cluster 2's variance are explained with transitivity and assortativity along dimension 3. As explored in Chapter 2, the characteristics that transitivity and assortativity capture can now be attributed to cluster 2, and therefore provide insight regarding the type and style of discussion that took place in these datasets.

Cluster 3 is unique in this study: it is the only cluster that contains one dataset. On closer inspection of its PC score in Table 4.4, however, it also has a substantially higher PC 2 score than any of the other datasets, as well as a high average PC 3 score. Assortativity contributes 43.34% to the second dimension and therefore, clearly defines cluster 3 as well as separating it from the other clusters, with some help from transitivity in PC 3, as seen in Figure 4.8.

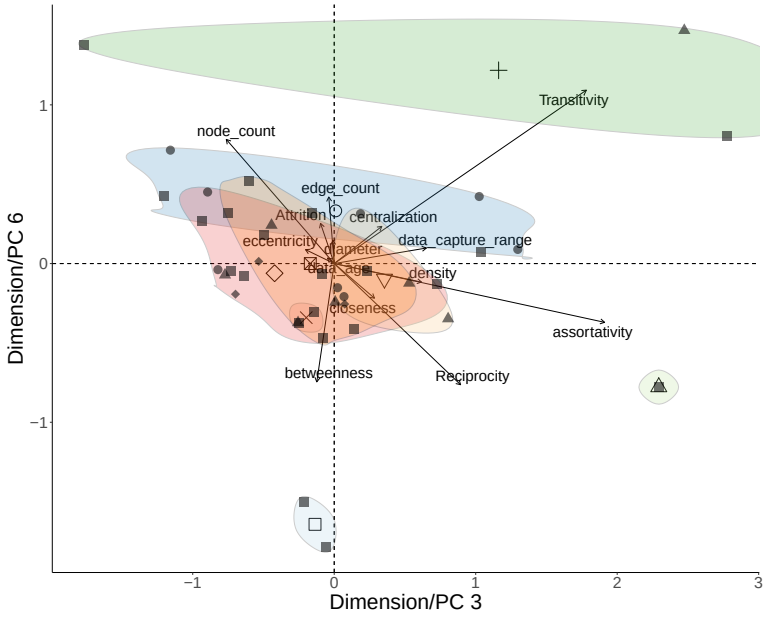
Cluster 4 also has a unique structure in this study: one of datasets is close to the centroid of the cluster, and the remaining two are relatively further from the centroid, as seen in Table 4.5. This is exemplified in Figures 4.5b and 4.7a, where the datasets for cluster 4 are spread far apart on dimension 1, 3 and 4. However, even though the PCA scores vary considerably, the cluster is still able to be clearly separated from the others, as seen in Figures 4.6a and 4.6b. This would suggest that the datasets in cluster 4 can be differentiated from the other datasets largely with PC 6, of which the positive contribution predominantly consists of transitivity and node count.

Cluster 5 consists of two datasets with the same categories and, as is evident in Table 4.5, both datasets have a substantially higher PC 1 score. This measure alone sets cluster 5 apart from the others, as is evident in Figure 4.5a. Table 4.4 in

Stellenbosch University https://scholar.sun.ac.za



(a) PC 1 & 3



(b) PC 3 & 6

Figure 4.7: Cluster 2: Bi-plot and Clusters

Cluster 2: Pair of biplots that indicate the variance in cluster 2 is largely explained with PC 3.

Stellenbosch University https://scholar.sun.ac.za

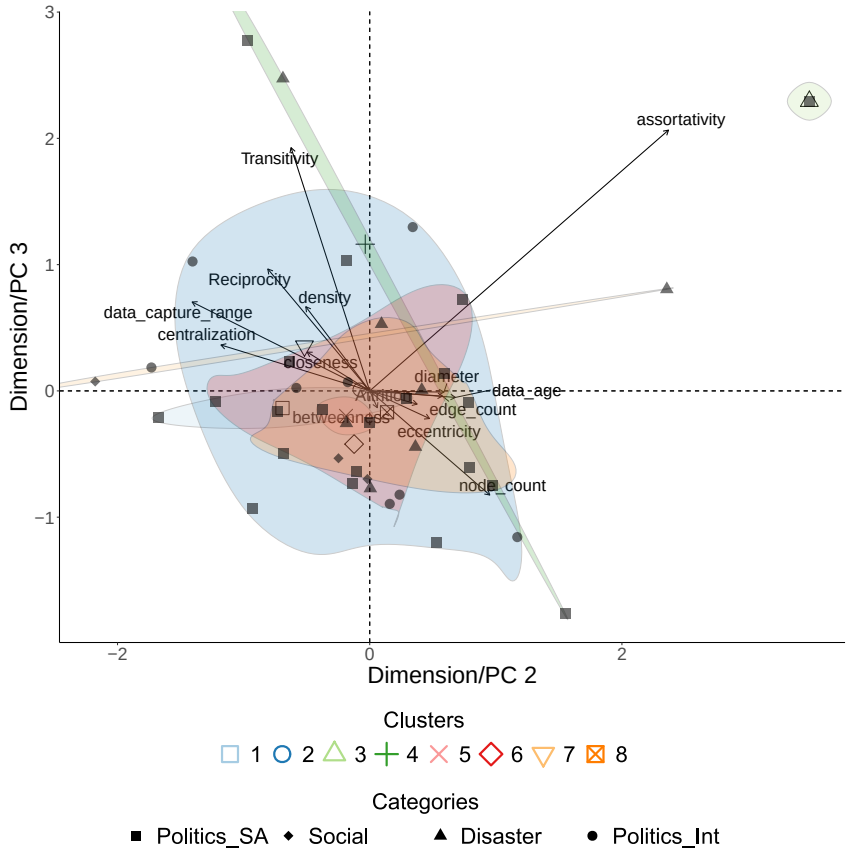


Figure 4.8: Cluster 3: Bi-plot and Clusters

Cluster 3: PC 2 and PC 3 clearly separate cluster 3 from the other clusters.

the previous section highlights the various measures that contribute the most to the first dimension; namely: closeness, density, betweenness, as well as the node and edge count descriptors. From Figure 4.2, it is clear that all of these measures contribute positively, along with centralisation, but to a lesser degree.

Although cluster 6 consists of predominantly South African political datasets, the most representative dataset contained disaster categories. Also, the mean euclidean distance for the dataset is 1.84, of which all but one—a Politics SA dataset—is less than one standard deviation (σ = 2.008) from the medoid. By inspecting Figure 4.9, it becomes evident that the majority of the variance in cluster 6 can be explained with dimension 5. For this dimension, the capture range contributes the majority (31.50%), with reciprocity and node count contributing 18.24% and 16% respectively, as seen in Table 4.4. This would suggest that the network structure of the datasets in the cluster is quite similar, except for the amount of days over which the data was collected, the node size and the reciprocity of the cluster. This is further supported by Table 4.5, where the outlying dataset has a higher PC 5 score than the rest, as well as in Figure 4.9.

Stellenbosch University https://scholar.sun.ac.za

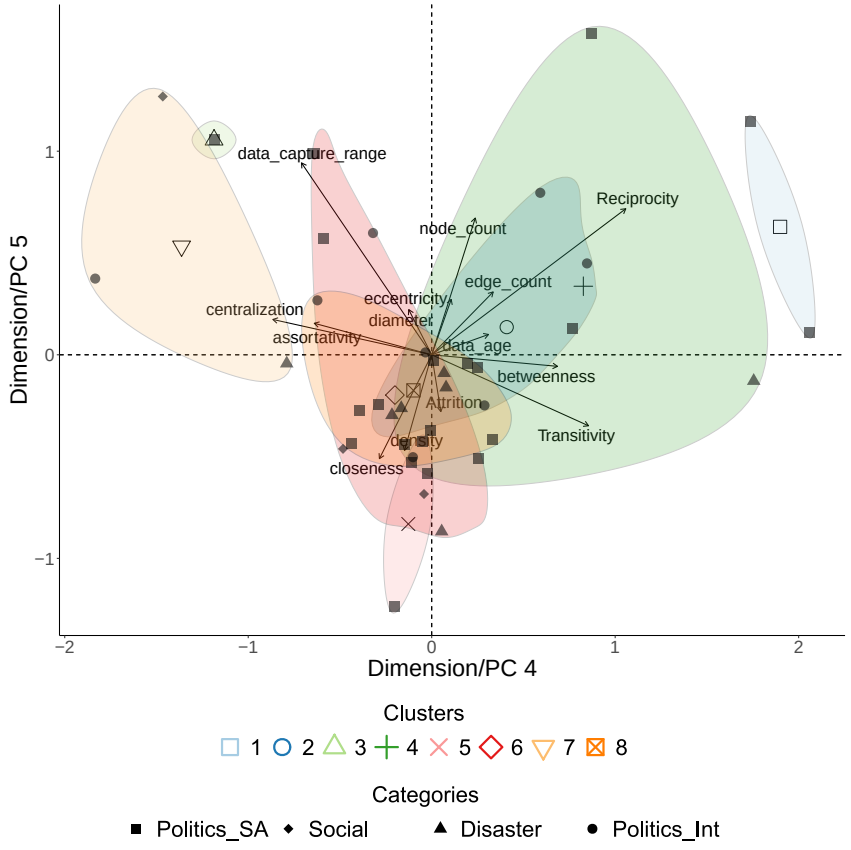


Figure 4.9: Cluster 6: Bi-plot and Clusters

Cluster 6: Most of the variance in cluster 6 is explained by dimension 5.

As seen in Table 4.5, cluster 7 has an assortment of categorised events, with the two politically aligned datasets a similar distance away from the centroid, and the disaster dataset with a distance of 4.64 away from the medoid. This is clearly shown in Figure 4.5b, where the political datasets have similar dimension three values. The datasets share a high PC 4 score, to which capture range and centralisation contributes positively, while the disaster has a higher PC 3 score, attributed to the transitivity and assortativity measures.

Cluster 8 is the largest cluster, with the lowest euclidean distance of 1.07. This suggests that all of these datasets share a similar network structure, as is evident by the low variance in all of their PCA scores, as seen in Table 4.5. The standard deviation of the euclidean distance is 0.57 for cluster 8. Given the highly representative medoid, only 0.01 away from the centre of the cluster, most of the other datasets in the cluster are at least one to two standard deviations away from the medoid. All of the biplots in Figures 1–8 show that cluster 8 is highly central for most dimensions. It is therefore probable that, due to the scaling of the biplots, cluster 8 is the average representation of all 39 datasets.

Stellenbosch University https://scholar.sun.ac.za

From this section, the possible insights that can be obtained based purely on the network structure of the datasets are exemplified, adding to the value proposed in this method. These findings illustrate the information that can be obtained from investigating the clustering results with no prior domain knowledge. The study proposes this approach as a universal application for any dataset to gain additional information. In the following section, however, one of the possible methods to analyse the results with minimal domain knowledge will be discussed.

4.4.2 Classification with Labelled Data

As discussed in Chapter 1, other classification studies in the literature performed the cluster analysis with the goal of classifying the social media data based on a label or ground truth. The goal of this thesis was not to provide a method to categorise events accurately based on pre-defined labels, but rather to prove the effectiveness of classifying social media data with network analysis metrics. Also, the labels, or rather categories, provided in this study were merely to contextualise the events around which the datasets were collected. Nevertheless, the study produced a crosstable to visualise the findings in the hypothetical case that the categories were used as labels. Table 4.6 lists the categories as fractions of datasets that fell in each of the four categories, with fractions of the total datasets in the final column and row.

Table 4.6: Crosstable for the Category and Clusters

Categories	**1** (5.1%)	**2** (15.4%)	**3** (2.6%)	**4** (7.7%)	**5** (5.1%)	**6** (25.6%)	**7** (7.7%)	**8** (30.7%)
Disaster (20.5%)	0	1	1	1	0	1	1	3
Politics Intl (17.9%)	0	0	0	2	2	0	1	2
Politics SA (53.8%)	1	4	0	0	0	8	1	7
Social (7.77%)	1	1	0	0	0	1	0	0

Note Crosstable between categories and clusters in number of total datasets that share a category. The percentage at categories indicates the proportion of datasets in each category and percentage under each cluster indicates the proportion of datasets in each cluster.

The most prominent finding from Table 4.6 is that the three largest clusters—2, 6 and 8—each contain three of the four possible categories. The largest cluster—8—has a similar percentage of datasets for each category: 37.5% of the disaster datasets, 33.3% of the SA Politics datasets and 28.6% of the International Politics datasets. With cluster 6, the SA politics contributes the most from it's category total, 38.1%. Cluster 2 has the highest percentage contributor of the social category. Notable outliers are cluster 3 and 5: each contain only one category of dataset. This suggests that other datasets classified using this model that share the network structure of cluster 3 and 5 would be classifiable. But, since these

Stellenbosch University https://scholar.sun.ac.za

clusters only consist of one and two datasets respectively, it reduces confidence in the results. The amount of SA Politics datasets in the study in comparison to the others skews the classification results and the lack of a clear majority category in any specific cluster indicates the need for a more granular and robust labelling of the data.

4.5 Chapter Summary

This chapter applied all the steps and methods outlined in Chapter 3. First, the descriptive statistics for all the datasets and measures were calculated and discussed to determine the necessity of the analysis options outlined in the previous chapter. Second, the analysis applied the network measures to the networked discussions and produced a quantitative characterisation of the network structure of each dataset. Third, PCA was performed on the results to obtain a set of scaled components that account for the correlation between all the network measures and data descriptors. Finally, the clustering analysis was performed. The quality and optimal number of clusters were determined to address the research question. The findings compared the results from the tests with the criteria outlined in Chapter 3 and found that social media discussions could be clustered successfully by utilising the network structure derived from interactions on Twitter. The possibility of using the primitive categories as labels was also explored to determine the viability of using the proposed method as a labelled classification. The results indicate the need for a more granular and robust labelling of the data to reach a reliable consensus.

Stellenbosch University https://scholar.sun.ac.za

Chapter 5

Discussion & Conclusion

5.1 Chapter Introduction

The purpose of this study was to determine the viability of clustering a variety of events that were discussed on Twitter by utilising interactions between users on the platform to create a network structure that could be measured. Social media discussions have become an increasingly common component of people's lives. Two examples will suffice here. They are used as references by mainstream media to determine the sequence of events; they are also used by companies to advertise and based on the online response, to analyse how effective those advertisements were (Jansen *et al.*, 2009, p. 2177). As discussed in Chapter 1, the problem of classifying social media data has been studied extensively in the database, machine learning, multi-media and text literature contexts. This study focused on exploring a social network approach to clustering discussions on Twitter and the benefits that this approach adds to utilising the interactions to create networks. This study sought to determine if it is possible to find useful data by applying unsupervised machine learning methods to unfiltered datasets to find underlying insights or patterns about the data in the form of clusters. The study proposed Social Network Analysis (SNA) as a field that can produce measures which calculate the variance between and in datasets and leads to meaningful insights.

In order to address the two research questions posed in Chapter 2 the theoretical background of SNA was explored. The aim was to understand events by mapping the relationships that connect the actors as a network. Various metrics that could be used to determine the characteristics of a network were explored. In this process nine network measures and five descriptors were explained and the differences they aim to highlight were elaborated upon. Chapter 3 set out the methodology for the analysis. In this chapter, the Twitter API was chosen to directly source the tweets from the Twitter database, and the datasets were formatted into networks upon which the measures could be applied. Chapter 3 also outlined the Principal Component Analysis approach that was used to account for high correlations between the measures as well as the clustering meth-

Stellenbosch University https://scholar.sun.ac.za

ods and tests used to produce and analyse the findings. Chapter 4 explored the output of the applied methodology and the choices performed to address the problem. In this final chapter, a broader discussion will be held on the implications of the clustering results, the network measures that were applied and the clustering of events with regards to the search parameters whereby each dataset was categorised. This chapter will conclude with the limitations of the study and recommendations for how the approach can be applied and improved in future work.

5.2 Discussion

As was shown in Chapter 4, the results from the Gap Statistic and Silhouette methods coincided with seven or eight clusters as the optimal number of clusters with a reasonable structure quality. The degree of success for the study was determined in Section 3.5.5, in which the study states that an ideal outcome exists where the gap statistic and silhouette methods results coincide within the requirements as a low number of optimal clusters and reasonable to high structure quality. In relation to the criteria specified by Kaufman and Rousseeuw (2013, p. 88), it was shown that by using network measures and descriptors, the approach can cluster social media data and achieve reasonable structure quality. Given the original number of 39 datasets, and maintaining a reasonable structure quality for the low number of optimal clusters, this implies that the research question was adequately satisfied. The evidence provided in the present study suggested that it is possible to cluster event-discussions successfully on social media by utilising their network structure as the means to classify the discussions without prior domain knowledge.

The research objective for the study was aimed towards proving the concept of network-based clustering as an effective method to uncover underlying patterns and subgroups in social media data. As found in Section 1.5, multiple studies have been performed on clustering social media data, towards the goal of classification, with a wide variety of academic and commercial applications. However, many previous studies did not consider a domain-agnostic approach to classifying social media data. And, in cases where network features were utilised, only a small fraction of the available network measures in the field was applied. The aim of this study was therefore to add value in a domain-agnostic approach that can be applied to a wide range of diverse events, by determining if networked social media discussions can be clustered. To address the second research question the following section will discuss the applicability of certain network measures to expand the literature on how they contribute to the case of a sample set of events.

Stellenbosch University https://scholar.sun.ac.za

5.2.1 Contribution of Individual Network Measure

By referring to Table 4.4 in Chapter 4, the contribution of individual network measures to the study as a whole can be discussed. Table 4.4 listed the 14 network measures and descriptors used in the analysis and outlined how much variance each measure and descriptor explained in a given principal component. By claiming that the approach proposed by this study was successful, the next step is to determine how much each measure contributed to the ability to cluster the datasets based on network measures. In addition, these results will be compared to the original discussion had in Chapter 2 as to why these measures were chosen, and their subsequent effect.

In Chapter 2, transitivity was described as a measure of the homophily of the networked discussions. By measuring transitivity, the study hoped to evaluate the structural balance and, to a degree, the measure proved to be effective. Structural balance refers to groups of people and effective relations who are substantively pleasing or lack interpersonal psychological tension. While it only accounted almost a 10th (8.44%) of the variance across the five principal components, the network measure proved instrumental in identifying political discussions, as seen in cluster 7, Figure 4.6a. Table 4.5 indicates that the relatively high score the disaster dataset obtained for PC 3, to which transitivity contributes 34.24%, separated it from the political datasets that had higher and similar scores for PC 4, to which reciprocity was the majority contributor.

Reciprocity is a measure that aimed to predict if events can be classified through clustering methods by determining the degree of social cohesion found in the discussion. While the measure contributed even less than transitivity overall, the measure contributed substantially to PC 4, 5 and 6 (see Table 4.4) which obtained high PCA scores for clusters 1, 4 and 7 (see Table 4.5) where political events were grouped with disaster or social events. This would suggest that discussions where more dialogues (i.e. bi-directional connections) were observed, is indicated by the reciprocity measure. It is important to note, however, that reciprocity shared a substantial contribution in PC 4 and 5 with the capture range descriptor. This suggests that the duration over which these discussions took place affected social cohesion.

The centrality based measures accounted for the majority of variance in the data, as shown in Table 4.4. Centralisation, closeness centrality, betweenness centrality and density are all measures that are dependent on the degree of nodes in a graph. These measures were the dominant contributors to the first principal component and accounted for over half of the variance in the data collectively. The measures also correlated with the network size. This confirms the theory discussed in Chapter 2: an actor with a large degree is in direct contact or is adjacent to many other actors. This was accounted for by the node and edge count descriptors. These measures were added to the study since they form a baseline of any social network's characteristics, and the applicability of these measures is further supported by the results.

Stellenbosch University https://scholar.sun.ac.za

Assortativity accounted for the majority of the variance in the second and third principal components and, as discussed in Chapter 2, aimed to indicate if the actors (nodes) in a network have a preference to connect with other actors that are similar, based on an similarity measure. As seen in Table 4.4, centralisation and transitivity were the second major network measure contributors in PC 2 and 3. This supports assortativity's results, since preference—or lack thereof—to connect with other actors correlates with positive or negative structural balance and the similarity measure is a node's degree. Assortativity produced high results for disaster datasets in cluster 3 and 7, as seen in Table 4.5, suggesting that a number of actors with similar degrees were prevalent in those discussions.

Of the remaining two network measures, neither contributed substantially to any of the principal components used in this study. In Chapter 2, diameter and eccentricity were included to provide measures for how sparse the network was. If there were any correlation between this, the longest geodesic and the centrality measures were applied in this study. In Table 4.2, these measures are shown to have a low standard deviation, further reinforcing the low variance in these results. This would mean that almost all datasets achieve similar scores on these measures. Since there was not much variation in these measures, they also did not contribute substantially to any of the principal components. The study recognises that these measures would be similar, since they are closely related (see Chapter 2), however it is interesting that they are similar across all datasets. A comparison of these measures applied to other forms of social media data would indicate if the observation is a result of the network structure of an online discussion or Twitter specifically. These measures did not influence the effectiveness of clustering social media data in any noteworthy capacity and, based on the evidence shown in this study, are unlikely to contribute to the classification of social media data.

Data age and attrition were two descriptors that also did not provide any substantial evidence for describing the variance among the datasets. Table 4.4 indicated that data age and attrition played a much smaller role than anticipated. It makes sense that the older the dataset is—as indicated in the start date column in Table 1—the more tweets are likely to be deleted. However, this result would indicate that the relationship between age and attrition is not linear. Also, while this is not the main focus of the analysis, the low contribution of attrition is noteworthy to research regarding bot detection on Twitter. As discussed in Chapter 3, Volkova and Bell (2017, p. 219) suggested that deleted accounts could also signal automated account activity, trolls and other manipulating agents in social media discussions. This result hints at a number of possible conclusions. For example, the idea that attrition rates are more related to data age than indications of bots.[1] Bots do not have a noteworthy effect on the network structure of discussions where they are prevalent. Similarly, attrition is a natural occurrence

[1] Bots or automated agents in this context are computer programs that utilise a programmatic method to interact on social networking sites (Ferrara *et al.*, 2016)

Stellenbosch University https://scholar.sun.ac.za

on Twitter and does not provide evidence of one specific characteristic of the network. Regarding the diameter and eccentricity network measures discussed previously, these measures do not account for substantial variance and therefore do not contribute much to the principal components. Further research with additional metrics will be necessary to determine if any of those suggestions have merit.

5.2.2 Classification of Diverse Events

The first research question posed in the study was addressed in Section 4.4: the requirement for clustering events, towards the goal of classification, by utilising the network structure derived from interactions on Twitter, without prior domain knowledge. Although this question has been satisfied, this does not prevent the clustering results from being analysed with domain knowledge and amplifying the value obtained from the proposed approach. Since the study chose not to analyse or gather any textual or user information, the only domain knowledge that the study retains is the various search parameters used in the data collection portion of the study. In the previous sections, the search parameters were simplified to broad categories that indicate the context within which the data was collected. This section will briefly explore the composition of the clusters, utilising the more detailed search parameters.

In the previous section, the clusters were analysed by the contributions of the various network measures to the high PCA scores found in each cluster. Similarly, the discussions around different events can follow the same structure even without an evident overlap in what was specifically discussed in the cluster. It is important to remember that the datasets are clustered based on their network structures. This would indicate that the content of these discussions cause them to share characteristics on a structural level despite apparent differences or similarities in content. For instance, cluster 1, which primarily consists of datasets with high betweenness and reciprocity values, was categorised as a South African (SA) political dataset and a social dataset. The SA political dataset revolved around a discussion regarding “telecoms”, “infrastructure”, “Fibre” and “mobile operator”, while the social dataset was gathered with the “sxsw”, “homeless”, “hotspot”, “Wi-Fi”, “4G network” search parameters. From these search parameters, the seemingly unrelated events are linked by the common discussion revolving around wireless networks and internet service providers.

Cluster 2 consisted of four SA political datasets with similar search parameters, the medoid of which contained “rhodesmustfall”, “protest” and “racism”. The other Politics SA datasets also contain “racism”, as well as “FeesMustFall”, “farm murders”, “expropriation” and “vicki momberg”—all terms associated with violent race protests in South Africa. The disaster dataset in this cluster contained terms related to violence, but not specifically race, such as “Ottawa shooting”, “gun violence” and “violent crime”. The datasets in this cluster were primarily classified based on similar values for transitivity and assortativity measures,

Stellenbosch University https://scholar.sun.ac.za

indicating participating actors are more likely to form connections with others similar to themselves. The social dataset revolved predominantly around the 2012 Superbowl sporting event. Although it does not refer to any violence related topics, the dataset was also the furthest away from the centroid, however, with a value of 7.32 for the euclidean distance. This suggests that an online discussion regarding violent crime and race would create the similar network structure, but due to presence of the social dataset, the results are not definitive. Further analysis of the specific actors, accounts, and interactions in these datasets, as well as topic modelling on the discussion, may reveal additional information.

Cluster 3 contained only a single dataset, regarding a disaster, with search parameters related to the Ebola crisis in 2014. This suggests that the discussion's network structure around this event was unique compared to other datasets used in this study. Recall that in Table 4.5 in Chapter 4, cluster 3 displayed higher PC 2 scores than any of the other datasets, in combination with a high PC 3 score resulted in a unique network structure that set the dataset apart from any of the other discussions. Since there are no other datasets in this cluster to compare it to, the network measures serve as the only insight into the clusters' characteristics.

Cluster 4 contained one international politics dataset with a euclidean distance of 1.95 from the centroid, and another a distance of 10.89 from the centroid. This would suggest that the first international political dataset—which contained the search parameters "irishge", "general election" and "voteie"—was the most representative of the cluster and revolved around the 2016 Irish general election. The other political dataset revolved around Brexit-related search terms, which links it with the former geographically. However, the disaster dataset in that cluster contained the search terms "Boston marathon" and "bombing" surrounding the tragic events of the 2013 Boston Marathon. These datasets were classified based on their similar transitivity and node count. Further analysis of the specific actors, accounts and interactions could indicate that the similar news media, political topics and responses caused high similarity measures for the transitivity metric and number of people involved in the discussion. Topic modelling on these datasets may reveal additional information that explains why they were placed in the same cluster.

The fifth cluster is much easier to explain with the search parameters associated with the cluster, since both international politics datasets revolved around political events in the United States of America. The first contained the terms relating to the 2012 presidential election and the second contained "panamapapers", which was a major political discussion in 2016 regarding leaked documents that detail financial and attorney-client information of off-shore entities. This would suggest that the same actors and interaction patterns would be involved, resulting in a similar network structure.

Cluster 6 is another SA politics dominated cluster in terms of the categories, but surprisingly, a disaster dataset is the most representative, with a low value of 0.05 for the euclidean distance. Inspecting the search parameters in more detail,

Stellenbosch University https://scholar.sun.ac.za

all the SA Political datasets pertain to violent race-centric protests which caused property destruction, with terms such as "FeesMustFall", "north west university", "university of free state", "farm killings" and "racism". The disaster dataset revolved around Hurricane Patricia in 2015, which also caused a substantial amount of property damage. The social dataset in this cluster revolved around the 2017 Super Bowl event, again with no clear indication as to why it would be related to the other datasets' content. The network structure of the social event is similar, as indicated by the similar PCA scores to the other values, with a value of 1.23 for the euclidean distance.

As explored in the previous subsection, cluster 7 contained a wide variety of diverse events. However, the two political datasets with a euclidean distance variance less than 0.52, and their search terms both contained "migrants" and "foreign national", suggests a related discussion around expatriates. These datasets shared a high PC 4 score, to which capture range and centralisation contributes positively. This suggests that a few popular actors either led the discussion or were the most discussed about. This is to be expected in a political discussion on polarising topics such as immigration.

In Chapter 4, Table 4.5, cluster 8 was found to have very low variance in the datasets' PCA scores and contains a wide variety of search parameters. The most representative dataset (distance = 0.01) revolved around a political discussion regarding Hong Kong in 2014, with the search parameters "H7N9", "bird flu", "protest" and more. This would suggest that the network structure for this event is centred around these topics, where all the datasets share similar centrality and assortativity results. For objective results, more detailed analysis of the tweet text and other metrics would be required.

With regards to Table 4.6 at the end of Chapter 4, the discussion in this section does provide some insight on how the categories could be better defined. For example, in Table 4.6, the SA politics category is split between cluster 6, 8 and 2, wherein the category is the majority contributor to cluster 2 and 6. This would suggest that the SA politics search parameter is mislabelled in cluster 8, since the most of the Disaster category actually fell in that cluster. These findings help further the discussion on how this approach could be used to detect underlying network patterns that could point out which events are structured differently than originally anticipated. The accuracy of this classification method would increase drastically with the addition of other non-network measures.

From this section, the practical applications of the clustering results have become evident. The method proposed in this study is effective in finding underlying patterns and similarities, but mainly serves to point researchers in the right direction. More detailed analysis is necessary for definite conclusions.

Stellenbosch University https://scholar.sun.ac.za

5.3 Limitations and Future Work

The limitations in this study were determined by two major considerations before the data collection and analysis began. *First*, the scope of the thesis was designed to solve the problem, specifically stated in Chapter 1, without domain knowledge. This limited the study from exploring a number of data collection options, as well as what data would be used once the tweet objects were obtained. As with any computational social science study, there are inherent limitations to controlling bias that stems from the data collection techniques and tools and the biases that may be present in the datasets used.

In Chapter 2, a number of social science data sourcing options were explored—namely questionnaires, surveys and interviews—but were rejected mainly due to the domain knowledge that would be required to understand the responses regarding a variety of diverse events. A computational method for data collection was decided upon instead and the Twitter API was chosen as the method with which to obtain the data. Chapter 3 explored further limitations to the amount and quality of data, with regards to API rate restrictions and data loss in the process of rehydration. The data, in the form of tweet objects, contained multiple options for obtaining measures with which to attempt the cluster analysis. Computational approaches to analysing the tweet text were explored, such as sentiment analysis and topic modelling. These were rejected for the same reason as the traditional data collection methods. Another reason supporting why these computational measures were not included, was the related work, outlined in Chapter 1, found that the problem of classifying social media data had been studied extensively: it had already been possible to classify social media data with these measures (Himelboim *et al.*, 2017; Ferrara *et al.*, 2013; Zhu *et al.*, 2014; Croitoru *et al.*, 2015). It is recommended that the result from this thesis be used in conjunction with the methods developed in those studies. Finally, the interaction data in the tweet object was chosen on which to perform SNA. To address the research question and to take ethical considerations into account, the interaction data was utilised to form networks of the discussions, and the unused components of the tweet object were discarded. This limited the study to performing the analysis only on the network measures obtained from the network. Therefore, the study succeeded in identifying underlying patterns and characterising events by utilising the network structure derived from interactions on Twitter, but additional measures will increase the granularity and reliability of the results.

Second, the study was limited by a number of technical challenges. A number of alternative approaches to PCA were available to perform dimensionality reduction in Section 3.5.2 as well as clustering in Section 3.5.3. The study applied Principal Component Analysis, since it is very widely used in a number of diverse disciplines (Jolliffe, 2002, p. 9). However, other more modern approaches to dimensionality reduction are available, such as t-distributed stochastic neighbour embedding (t-SNE) and Uniform manifold approximation and projection

Stellenbosch University https://scholar.sun.ac.za

(UMAP). It is important to note that the purpose of this study was a proof of concept: that future work is recommended to use multiple methods for each stage of this process. Technical limitations—such as the amount and non-normal distribution of the data—prevented this study from being performed multiple times with different methods for each stage and still be completed in the time allotted for the study. Furthermore, methods such as UMAP also assume that the data is uniformly distributed. According to Table 4.2, the data is highly skewed. With regards to clustering, another widely used approach is hierarchical clustering. Agglomerative Hierarchical clustering (AGNES) is a primitive form of clustering. It starts with each observation in singleton clusters and then sequentially merges clusters based on some "Linkage" method (Liu *et al.*, 2011, p. 147). This approach suffers from reliability issues when subjected to noisy or real-world data. When clusters are merged/split, the decision is permanent, and the dendrogram plots can be misinterpreted (Liu *et al.*, 2011, p. 150). K-means has been proven to provide reliable results, given that the optimal value for k was chosen, as determined by the gap statistic and silhouette methods (Liu *et al.*, 2011, p. 140). Similarly to the clustering techniques, more network measures can also be applied in future classification studies, for example centralization can be divided into in-degree and out-degree centralization, as hinted at in Chapter 2. This is acknowledged as a technical limitation in terms of computation time and also the scope available for this study.

Finally, with regards to the data sources utilised in this study, the results of this study are limited to the current interaction patterns offered by Twitter. It is currently unknown whether the findings would generalise, firstly, to other interaction forms and, secondly, to other social media platforms. The affordances of other platforms (e.g. Instagram and Facebook) enable different interaction forms. While the study has shown that network measures (specifically the seven network measures and three descriptors that contributed substantially to the principal components) enable the successful clustering of event-discussions on Twitter, future research is needed to determine the extent to which this outcome holds across other popular social media platforms.

It is important to note that the purpose of this study was a proof of concept. Future work is recommended to use multiple methods for each stage of this process. Moreover, the approach proposed by this thesis recommends that future work incorporate additional measures—such as sentiment analysis, topic modelling and other Twitter user features—to increase the reliability and value of the results.

5.4 Conclusion

The goal of the study was to explore computationally the clustering of events, as discussed on social media, by using the interactions as a means to create a network whereby the various discussions can be characterised. The motivation was

Stellenbosch University https://scholar.sun.ac.za

to increase the understanding of how much each of the measures contributed to differentiating between the discussions and how useful they were in this context. This study was a first step in clustering social media data using a wide range of network measures and data descriptors, to maximise the value obtained by analysing social media data.

The study achieved this by exploring a variety of data options for addressing the research question. Traditional approaches were considered, and computational methods were researched and found to have been extensively applied to non-network features for clustering purposes. Therefore, SNA was chosen as the method whereby to approach the problem. By applying in-depth knowledge, obtained from investigating Twitter's database structure, a novel approach to classifying interactions was designed and networks were generated from these interactions. A wide array of network measures were explored to quantify the network structure characteristics of the datasets. In addition, a series of data descriptors were obtained to account for variations in the data caused by the collection method and data descriptives. In order to expand the literature on what they define in the context of this study, the network measures and data descriptors were subjected to dimensionality reduction to account for co-variance in the measurements and to evaluate the contribution of each network measure. The resulting principal components were used to cluster the discussions of diverse events, by applying clustering methods. The quality and quantity of those clusters were evaluated. A set of criteria for the classification of the clustering results with regards to quantity and quality was determined.

The study found that the approach produced an optimal number of clusters with reasonable structure quality without requiring any domain knowledge to produce them. The method proposed in this study is effective in finding underlying patterns and similarities, but mainly serves to point researchers in the right direction: more detailed analysis is necessary for definite conclusions and labelled categorisation. The study recognises the prior work performed in clustering social media data and recommends that future work include a wide variety of user features, sentiment, topic, and network measures. Furthermore, the study can be expanded upon by testing alternative dimensionality reduction and clustering methods at each stage of the proposed approach. The study furthered the understanding of clustering social media data by utilising the social network analysis approach and the various network measures and data descriptors that were discussed.

Stellenbosch University https://scholar.sun.ac.za

List of References

Abdullah, N.A., Nishioka, D. and Murayama, Y. (2016). Questionnaire testing: Identifying Twitter user's information sharing behavior during disasters. *Journal of Information Processing*, vol. 24, no. 1, pp. 20–28. ISSN 18826652.

Acar, A. and Muraki, Y. (2011). Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, vol. 7, no. 3, pp. 392–402. ISSN 14778394.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. (2011). Sentiment Analysis of Twitter Data. *Lecture Notes in Computer Science*, vol. 8883, pp. 33–52. ISSN 16113349.

Alshehri, A. and Keefe, R.M.O. (2018). Using the Twitter Platform as a Research Method in the Information and Social Media Age. In: *Mediterranean Conference on Information Systems (MCIS)*, pp. 1–15. Corfu, Greece.

Amaro, S., Duarte, P. and Henriques, C. (2016). Travelers' use of social media: A clustering approach. *Annals of Tourism Research*, vol. 59, pp. 1–15. ISSN 01607383.

Ashktorab, Z., Brown, C., Nandi, M. and Culotta, A. (2014). Tweedr: Mining Twitter to Inform Disaster Response. In: *Proceedings of the 11th International ISCRAM Conference*, May. University Park, Pennsylvania, USA.
Available at: `http://tweedr.dssg.io`

Babbie, E. (2010). *The Basics of Social Research*. 5th edn. Wadsworth, Belmont. ISBN 9780495812241.

Blake, R.R. and Moreno, J.L. (1954). *Who Shall Survive?* Nervous and Mental Disease Publishing Co., Washington, DC, USA.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3, pp. 993–1022. ISSN 15324435.

Boyd, D.M. and Ellison, N.B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230. ISSN 10836101.

Bradley, A. and James, R.J. (2019). How are major gambling brands using Twitter? *International Gambling Studies*, vol. 19, no. 3, pp. 451–470. ISSN 14794276.
Available at: `https://doi.org/10.1080/14459795.2019.1606927`

Stellenbosch University https://scholar.sun.ac.za

Burt, R.S. (2005). *Brokerage & Closure. An Introduction to Social Capital*. Oxford University Press, New York, New York, USA. ISBN 9780199249145.

Buscema, M., Ferilli, G., Massini, G. and Zavarrone, E. (2015). Media Content analysis on Online Hate Speech. Tech. Rep., Coalition of Positive Messengers to Counter Online Hate Speech.

Cai, H., Yang, Y., Li, X. and Huang, Z. (2015). What are popular: Exploring twitter features for event detection, tracking and visualization. In: *Proceedings of the 2015 ACM Multimedia Conference*, pp. 89–98. Brisbane, Australia. ISBN 9781450334594.

Cameron, M.P., Barrett, P. and Stewardson, B. (2016). Can Social Media Predict Election Results? Evidence From New Zealand. *Journal of Political Marketing*, vol. 15, no. 4, pp. 416–432. ISSN 15377865.
Available at: `http://dx.doi.org/10.1080/15377857.2014.959690`

Cattell, R.B. (1966). Multivariate Behavioral Translator disclaimer The Scree Test For The Number Of Factors. *Multivariate behavioral research*, vol. 1, no. 2, pp. 245–276.

Cornelissen, L.A. (2013). *IDENTITY POSITIONING FOR TRUST: A Narrative Analysis on Consultant Identity Construction (MA Thesis), Stellenbosch University*. Ph.D. thesis, Stellenbosch University.

Cornelissen, L.A., Theron, P.P., De Bruyn, C., Schoonwinkel, P., Ledingwane, M.K. and Barnett, R.J. (2019). Cross-sample community detection and sentiment analysis: South African twitter. In: *ACM International Conference Proceeding Series*. Association for Computing Machinery, Skukuza, South Africa. ISBN 9781450372657.

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J. and Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems*, vol. 53, pp. 47–64. ISSN 01989715.
Available at: `http://dx.doi.org/10.1016/j.compenvurbsys.2014.11.002`

Csardi, G. and Tamas, N. (2006). The igraph software package for complex network research.
Available at: `http://igraph.org`

Cui, R., Gallino, S., Moreno, A. and Zhang, D.J. (2018). The Operational Value of Social Media Information. *Production and Operations Management*, vol. 27, no. 10, pp. 1749–1769. ISSN 19375956.

Delamater, J. (2006). *Handbook of Social Psychology*. 2nd edn. Springer, New York, New York, USA. ISBN 9789400767713.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001). *Pattern Classification*. 2nd edn. John Wiley & Sons, Inc., New York, New York, USA. ISBN 9781118586006.

Ferrara, E., JafariAsbagh, M., Varol, O., Qazvinian, V., Menczer, F. and Flammini, A. (2013). Clustering memes in social media. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013*,

Stellenbosch University https://scholar.sun.ac.za

pp. 548–555. Association for Computing Machinery, Niagra, Ontario, Canada. ISBN 9781450322409.

Ferrara, E., Varol, O., Davis, C.A., Menczer, F. and Flammini, A. (2016 6). The Rise of Social Bots. *Communications of the ACM*, vol. 59, no. 7, pp. 96–104. ISSN 00010782. Available at: `http://dl.acm.org/citation.cfm?doid=2963119.2818717`

Freeman, L.C. (1979). Centrality in social networks. *Social Networks*, vol. 1, no. 3, pp. 215–239. ISSN 03788733.

Frenz, V. (2019). *Cognitive Structural Accuracy (MA Thesis), Stellenbosch University*. Ph.D. thesis, Stellenbosch University.

Gaál, Z., Szabó, L., Obermayer-Kovács, N. and Csepregi, A. (2015). Exploring the Role of Social Media in Knowledge Sharing. *The Electronic Journal of Knowledge Management*, vol. 13, no. 3, pp. 185–194. ISSN 1479-4411.

Gerlach, M., Peixoto, T.P. and Altmann, E.G. (2018). A network approach to topic models. *Science Advances*, vol. 4, no. 7. ISSN 23752548.

Golbeck, J. (2013). *Analyzing the Social Web*. Morgan Kaufmann Publishers Inc., Waltham, Massachusetts, USA. ISBN 9780124055315.

Goncalves, P., Aradujo, M., Benevenuto, F. and Meeyoung, C. (2013). Comparing and Combining Sentiment Analysis Methods. In: *Proceedings of the First ACM Conference on Online Social Networks*, pp. 27–37. Boston, Massachusetts, USA.

Granovetter, M.S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380.

Harary, F. (1994). *Graph Theory*. CRC Press, Taylor & Francis Group, Boca Raton, Florida, USA. ISBN 9780201410334.

Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, vol. 30, no. 2, pp. 87–93. ISSN 0137477X.

Himelboim, I., Smith, M.A., Rainie, L., Shneiderman, B. and Espina, C. (2017). Classifying Twitter Topic-Networks Using Social Network Analysis. In: *Social Media and Society*, vol. 3, pp. 1–13. Sage, Chicago, Illinois, USA. ISSN 20563051.

Hodas, N.O. and Lerman, K. (2014). The simple rules of social contagion. Tech. Rep., USC Information Sciences Institute, Marina del Rey, California, USA.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Jersey, USA. ISBN 0-13-022278-X.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, New York, New York, USA. ISBN 9781461471370.

Stellenbosch University https://scholar.sun.ac.za

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2169–2188. ISSN 15322882.

Jolliffe, I.T. (2002). *Principal Component Analysis, Second Edition*, vol. 30. 2nd edn. Springer, Aberdeen, United Kingdom. ISBN 0387954422.
Available at: `http://onlinelibrary.wiley.com/doi/10.1002/0470013192.bsa501/full`

Kaufman, L. and Rousseeuw, P.J. (2013). *Finding Groups in Data: An Introduction to Cluster Analysis*, vol. 3. John Wiley & Sons, Inc., Hoboken, New Jersey, USA. ISBN 0-471-73578-7.

Ledford, B.H. (2020 6). Computing humanity. *Nature*, vol. 582, no. 7812, pp. 328–330.
Available at: `https://www.nature.com/articles/d41586-020-01747-1`

Liu, G., Wang, Y. and Orgun, M.A. (2011). Trust transitivity in complex social networks. In: *Proceedings of the National Conference on Artificial Intelligence*, vol. 2, pp. 1222–1229. San Francisco, California, USA. ISBN 9781577355090.

McPherson, M., Smith-lovin, L. and Cook, J.M. (2001). BIRDS OF A FEATHER: Homophily in Social Networks. *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444.

Morstatter, F., Wu, L., Nazer, T.H., Carley, K.M. and Liu, H. (2016). A New Approach to Bot Detection: Striking the Balance Between Precision and Recall. In: *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 533–540. San Francisco, California, USA. ISBN 9781509028467.
Available at: `https://ieeexplore.ieee.org/document/7752287/`

Newman, M.E. (2002). Assortative Mixing in Networks. *Physical Review Letters*, vol. 89, no. 20, pp. 1–5. ISSN 10797114.

Paulussen, S. and Harder, R.A. (2014). Social Media References in Newspapers. *Journalism Practice*, vol. 8, no. 5, pp. 542–551. ISSN 1751-2786.

Pitts, F.R. (1978). The medieval river trade network of Russia revisited. *Social Networks*, vol. 1, no. 3, pp. 285–292. ISSN 03788733.

Popescu, A.M. and Pennacchiotti, M. (2010). Detecting controversial events from twitter. In: *International Conference on Information and Knowledge Management, Proceedings*, pp. 1873–1876. Toronto, Ontario, Canada. ISBN 9781450300995.

R Core Team (2020). R: A Language and Environment for Statistical Computing.
Available at: `https://www.r-project.org/`

Reale, E., Avramov, D., Canhial, K., Donovan, C., Flecha, R., Holm, P., Larkin, C., Lepori, B., Mosoni-Fried, J., Oliver, E., Primeri, E., Puigvert, L., Scharnhorst, A., Schubert, A., Soler, M., Soòs, S., Sordé, T., Travis, C. and Van Horik, R. (2018). A review of literature on evaluating the scientific, social and political impact of social sciences and humanities research. *Research Evaluation*, vol. 27, no. 4, pp. 298–308. ISSN 14715449.

Stellenbosch University https://scholar.sun.ac.za

Robins, G. (2015). *Doing Social Network Research*, vol. 21. Sage, London, United Kingdom. ISBN 9780132478663.
Available at: `https://www.faa.gov/data_research/aviation/aerospace_forecasts/media/FY2017-37_FAA_Aerospace_Forecast.pdf`

Robson, C. and McCartan, K. (2013). *Real World Research: A Resource for Users of Social Research Methods in Applied Settings*. 4th edn. John Wiley & Sons, Inc., Chichester, West Sussex, United Kingdom. ISBN 9781119144854.

Rosen, A. and Ihara, I. (2017). Giving you more characters to express yourself.
Available at: `blog.twitter.com/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself`

Sanandres, E., Madariaga, C. and Abello, R. (2018). Topic modeling of Twitter conversations. In: *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, June. UniversItalia, Rome, Italy. ISBN 9788832931372.

Sayce, D. (2020). The Number of tweets per day in 2020.
Available at: `https://www.dsayce.com/social-media/tweets-day/#:~:text=Numberoftweetsperday2016,August2014with661million.`

Scott, J. (2011). Social network analysis: developments, advances, and prospects. *Social Network Analysis and Mining*, vol. 1, no. 1, pp. 21–26. ISSN 18695469.

Shaham, E., Yu, H. and Li, X.L. (2016). On finding the maximum edge biclique in a bipartite graph: A subspace clustering approach. In: *16th SIAM International Conference on Data Mining 2016, SDM 2016*, pp. 315–323. Miami, Florida, USA. ISBN 9781510828117.

Shamma, D. and Kennedy, L. (2010). Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events? In: *Proc. ACM Conf. on Computer-Supported Cooperative Work*, pp. 589–593. Savannah, Georgia. ISBN 9781605587950.
Available at: `http://www.research.yahoo.net/files/horizon4s-shamma.pdf`

Shimbel, A. (1953). Structural parameters of communication networks. *The Bulletin of Mathematical Biophysics*, vol. 15, no. 4, pp. 501–507. ISSN 00074985.

Sigala, M. and Chalkiti, K. (2015). Knowledge management, social media and employee creativity. *International Journal of Hospitality Management*, vol. 45, pp. 44–58. ISSN 02784319.
Available at: `http://dx.doi.org/10.1016/j.ijhm.2014.11.003`

Squartini, T., Picciolo, F., Ruzzenenti, F. and Garlaschelli, D. (2013). Reciprocity of weighted networks. *Scientific Reports*, vol. 3:2729. ISSN 20452322.

Stefanidis, A., Crooks, A. and Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, vol. 78, no. 2, pp. 319–338. ISSN 03432521.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society: Series B*, vol. 63, no. Part 2, pp. 411–423.

Stellenbosch University https://scholar.sun.ac.za

Twitter (2019). Selected Company Metrics and Financials. Tech. Rep. 1. Available at: `s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Selected-Company-Metrics-and-Financials.pdf`

Volkova, S. and Bell, E. (2017). Identifying effective signals to predict deleted and suspended accounts on Twitter across languages. In: *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, Icwsm, pp. 290–298. Montreal, Quebec, Canada. ISBN 9781577357889.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. 1st edn. Cambridge University Press, New York, New York, USA. ISBN 0521382696.

Zhang, Z., He, Q. and Zhu, S. (2017). Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method. *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 396–414. ISSN 0968090X.

Zhu, L., Galstyan, A., Cheng, J. and Lerman, K. (2014). Tripartite graph clustering for dynamic sentiment analysis on social media. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1531–1542. Association for Computing Machinery, Snowbird, Utah, USA. ISBN 9781450323765. ISSN 07308078.

Zubiaga, A. (2018). A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, vol. 69, no. 8, pp. 974–984. ISSN 23301643.

Stellenbosch University https://scholar.sun.ac.za

Appendices

Stellenbosch University https://scholar.sun.ac.za

Additional Figures and Tables

Stellenbosch University https://scholar.sun.ac.za

Table 1: Descriptive Statistics of the 39 Datasets

Category	Ogri. Count	Count	Attrition	Users	Date Range	Start Date	End Date
Politics_Int	5044378	1122889	78	680459	30	2016-04-03	2016-05-03
Disaster	3430386	914091	73	1246776	20	2013-04-15	2013-04-16
Politics_Int	1826289	561575	69	196474	69	2016-02-24	2016-05-03
Politics_Int	1743152	407596	77	388007	83	2015-09-02	2015-11-24
Politics_Int	1740257	546850	69	555034	3	2012-11-05	2012-11-08
Social	1659474	572006	66	601241	4	2012-02-03	2012-02-07
Social	1563447	703586	55	321651	14	2012-03-08	2012-03-22
Politics_Int	1524165	444715	71	321692	3	2014-09-17	2014-09-20
Politics_Int	1188371	440653	63	196241	25	2014-09-26	2014-10-20
Disaster	1151219	297233	74	361211	45	2015-10-24	2015-12-08
Disaster	1149252	319088	72	340197	3	2016-03-27	2016-03-30
Disaster	1075863	324950	70	340197	3	2016-03-27	2016-03-30
Disaster	1007866	364710	64	177578	11	2016-04-17	2016-04-28
Disaster	986524	359630	64	246441	31	2014-07-01	2014-07-31
Politics_SA	925912	396554	57	126160	13	2017-12-10	2017-12-23
Politics_Int	758802	313216	59	54076	32	2016-02-03	2016-03-06
Disaster	702585	208498	70	207512	1	2016-03-29	2016-03-30
Politics_SA	508608	179464	65	64509	18	2016-09-19	2016-10-07
Disaster	264625	87634	67	68394	6	2014-12-05	2014-12-11
Politics_SA	168529	56958	66	65662	17	2017-02-24	2017-03-13
Politics_SA	113878	62193	45	28219	2	2017-03-16	2017-03-17
Politics_SA	105150	94178	10	30679	98	2017-11-20	2018-02-26
Politics_SA	105150	98244	7	30685	98	2017-11-20	2018-02-26
Social	99998	21529	78	41077	31	2017-01-12	2017-02-12
Politics_SA	99993	30198	70	32546	11	2015-10-15	2015-10-26
Politics_SA	93802	54473	42	31019	24	2017-05-22	2017-06-15
Politics_SA	77786	29741	62	25299	36	2015-03-19	2015-04-24
Politics_SA	74451	39272	47	20677	16	2017-01-25	2017-02-10
Politics_SA	71851	39788	45	23944	67	2018-03-27	2018-06-02
Politics_SA	69435	29067	58	21360	84	2018-04-20	2018-07-13
Politics_SA	60586	28612	53	30462	37	2018-01-24	2018-03-02
Politics_SA	56444	35920	36	19609	57	2018-06-07	2018-08-03
Politics_SA	51116	12928	75	30022	42	2018-03-28	2018-05-09
Politics_SA	49998	15616	69	17630	18	2017-10-30	2017-10-31
Politics_SA	48064	16290	66	12656	2	2016-02-23	2016-02-25
Politics_SA	43196	13964	68	16101	11	2017-10-25	2017-11-05
Politics_SA	41597	15511	63	27885	1	2016-07-18	2016-07-19
Politics_SA	28977	11727	60	10919	26	2018-04-13	2018-05-09
Politics_SA	17238	8465	51	8076	13	2018-06-07	2018-06-20

Stellenbosch University https://scholar.sun.ac.za

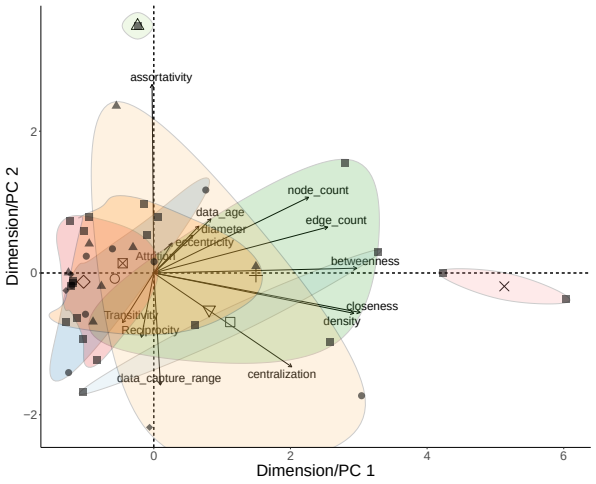
Table 2: Un-scaled Descriptive Statistics for the 14 Network Measures and Descriptors.

Measure	*Mean*	*Std dev*	*Median*	*Min*	*Max*	*Skew*	*Kurtosis*
Trans	$5.75{\cdot}10^{-3}$	$7.55{\cdot}10^{-3}$	$3.64{\cdot}10^{-3}$	$1.34{\cdot}10^{-4}$	$3.71{\cdot}10^{-2}$	2.44	6.24
Reci	$3.45{\cdot}10^{-2}$	$3.36{\cdot}10^{-2}$	$2.62{\cdot}10^{-2}$	$5.52{\cdot}10^{-4}$	$1.86{\cdot}10^{-1}$	2.53	8.17
Centr	$4.36{\cdot}10^{-3}$	$5.77{\cdot}10^{-3}$	$1.66{\cdot}10^{-3}$	$2.98{\cdot}10^{-4}$	$2.29{\cdot}10^{-2}$	1.78	2.19
Close	$1.29{\cdot}10^{-7}$	$2.34{\cdot}10^{-7}$	$3.42{\cdot}10^{-8}$	$1.36{\cdot}10^{-9}$	$9.42{\cdot}10^{-7}$	2.42	4.89
Dens	$1.93{\cdot}10^{-7}$	$3.24{\cdot}10^{-7}$	$6.05{\cdot}10^{-8}$	$8.23{\cdot}10^{-9}$	$1.15{\cdot}10^{-6}$	2.03	2.77
Assort	$4.09{\cdot}10^{-2}$	$1.91{\cdot}10^{-1}$	$-1.71{\cdot}10^{-3}$	$-2.27{\cdot}10^{-1}$	$8.89{\cdot}10^{-1}$	2.51	8.16
Betw	$2.86{\cdot}10^{-4}$	$6.54{\cdot}10^{-4}$	$2.17{\cdot}10^{-5}$	$2.12{\cdot}10^{-7}$	$2.80{\cdot}10^{-3}$	2.71	6.72
Diam	$1.89{\cdot}10^{1}$	6.98	$1.90{\cdot}10^{1}$	3.00	$3.80{\cdot}10^{1}$	$3.80{\cdot}10^{-1}$	$5.43{\cdot}10^{-1}$
Ecc	$1.43{\cdot}10^{1}$	4.59	$1.30{\cdot}10^{1}$	7.00	$2.50{\cdot}10^{1}$	$3.09{\cdot}10^{-1}$	$-8.25{\cdot}10^{-1}$
Range	$2.83{\cdot}10^{1}$	$2.77{\cdot}10^{1}$	$1.78{\cdot}10^{1}$	1.22	$9.76{\cdot}10^{1}$	1.17	$3.04{\cdot}10^{-1}$
Node	$1.80{\cdot}10^{5}$	$2.51{\cdot}10^{5}$	$5.41{\cdot}10^{4}$	$8.08{\cdot}10^{3}$	$1.25{\cdot}10^{6}$	2.29	6.17
Edge	$2.38{\cdot}10^{5}$	$2.72{\cdot}10^{5}$	$9.42{\cdot}10^{4}$	$8.46{\cdot}10^{3}$	$1.12{\cdot}10^{6}$	1.34	1.40
Age	$1.47{\cdot}10^{3}$	$6.32{\cdot}10^{2}$	$1.50{\cdot}10^{3}$	$7.00{\cdot}10^{2}$	$3.04{\cdot}10^{3}$	$8.56{\cdot}10^{-1}$	$3.90{\cdot}10^{-2}$
Attr	$6.04{\cdot}10^{1}$	$1.59{\cdot}10^{1}$	$6.50{\cdot}10^{1}$	7.00	$7.80{\cdot}10^{1}$	−1.77	3.24

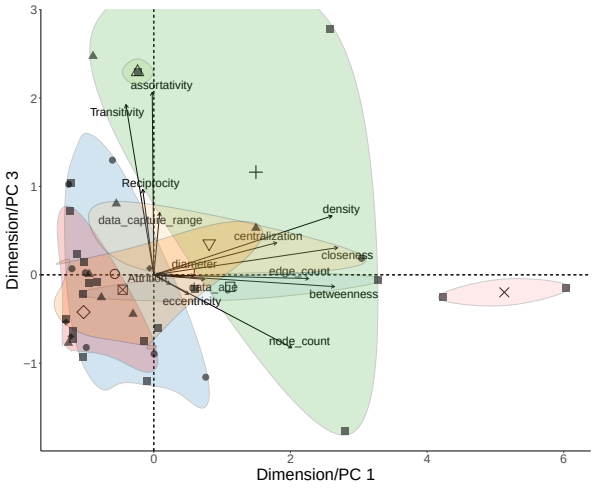
Note: Summary Statistics of the measures, unscaled.

Stellenbosch University https://scholar.sun.ac.za

Figure 1: Meta-figure 1 for Cluster Results



(a) PC 1 & 2

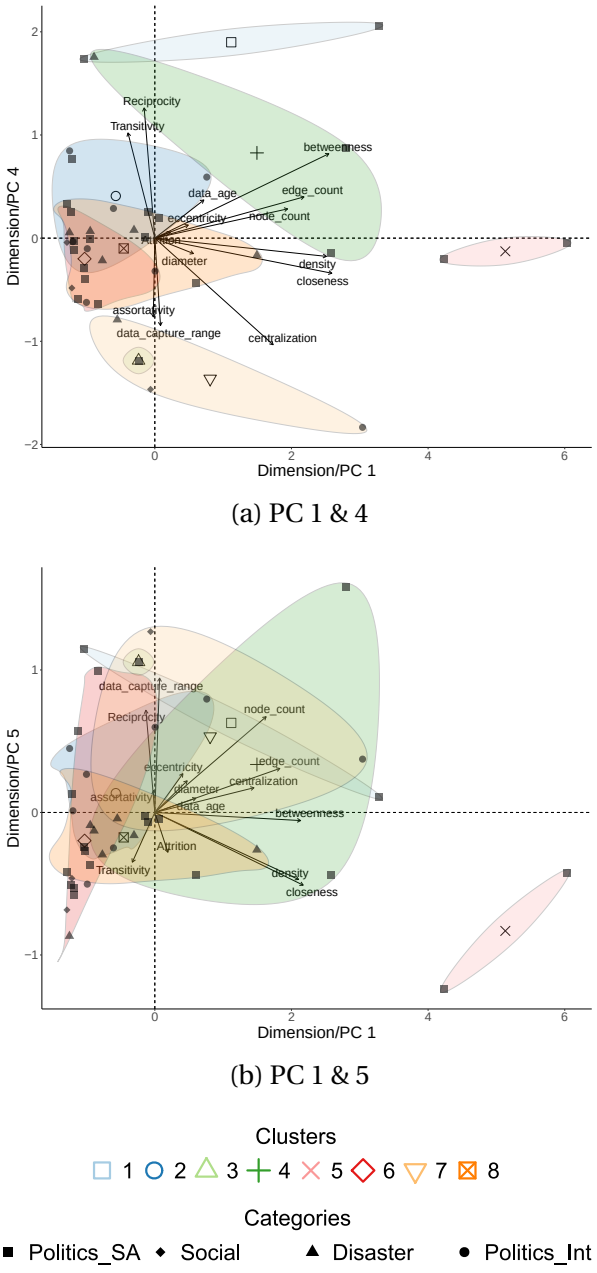


(b) PC 1 & 3

Note: The categories of the data sets visualised on pairs of principal components

Stellenbosch University https://scholar.sun.ac.za

Figure 2: Meta-figure 2 for Cluster Results



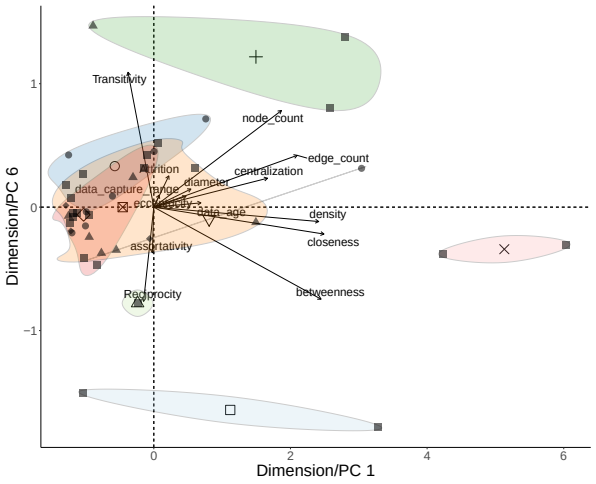
(a) PC 1 & 4

(b) PC 1 & 5

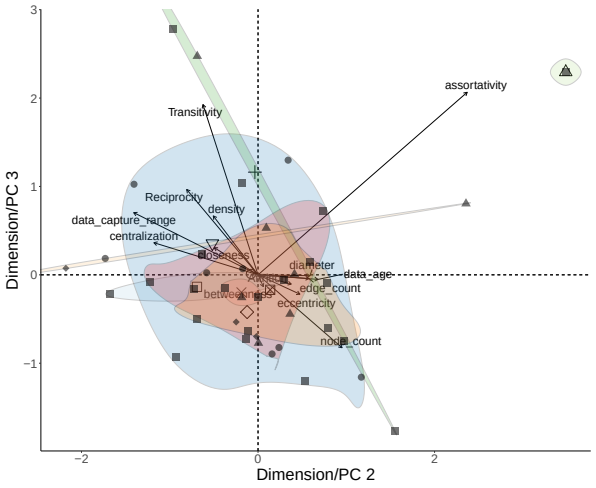
Note: The categories of the data sets visualised on pairs of principal components

Stellenbosch University https://scholar.sun.ac.za

Figure 3: Meta-figure 3 for Cluster Results



(a) PC 1 & 6

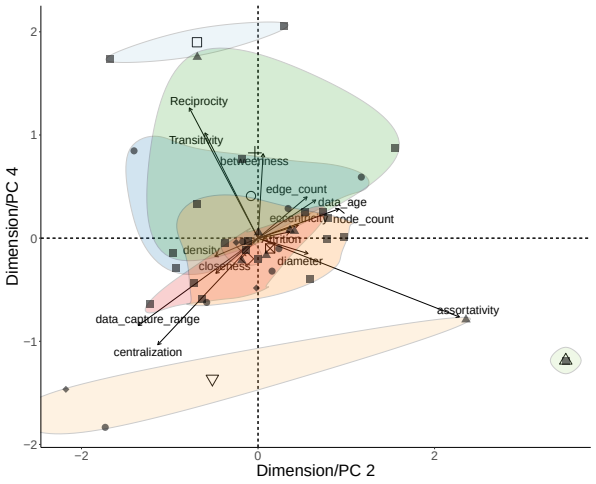


(b) PC 2 & 3

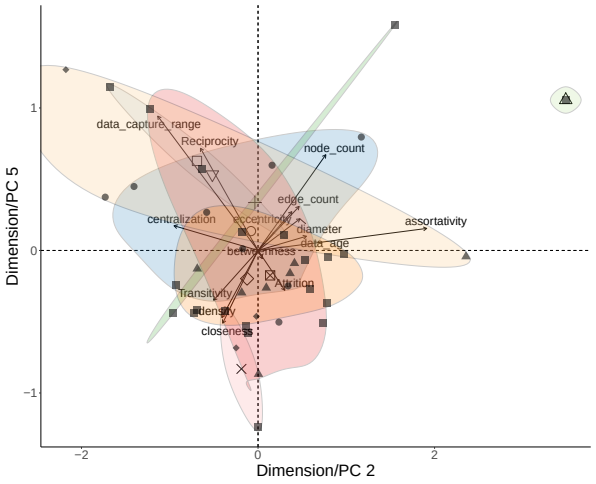
Note: The categories of the data sets visualised on pairs of principal components

Stellenbosch University https://scholar.sun.ac.za

Figure 4: Meta-figure 4 for Cluster Results



(a) PC 2 & 4

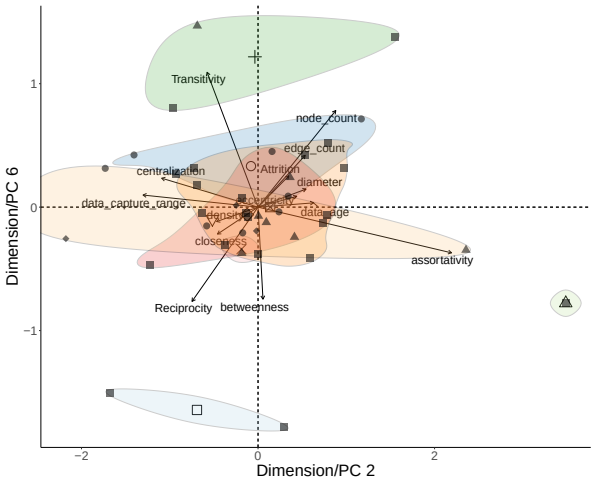


(b) PC 2 & 5

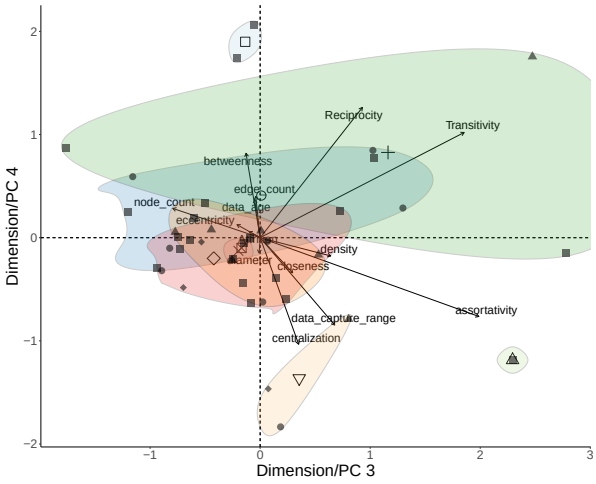
Note: The categories of the data sets visualised on pairs of principal components

Stellenbosch University https://scholar.sun.ac.za

Figure 5: Meta-figure 5 for Cluster Results



(a) PC 2 & 6

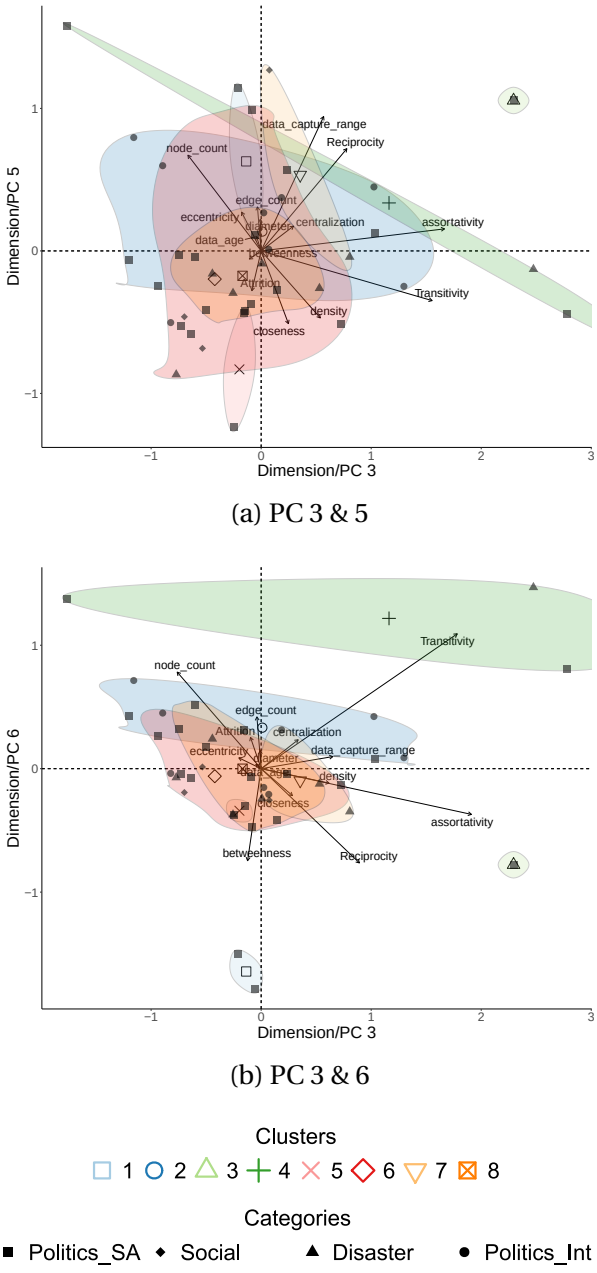


(b) PC 3 & 4

Note: The categories of the data sets visualised on pairs of principal components

Stellenbosch University https://scholar.sun.ac.za

Figure 6: Meta-figure 6 for Cluster Results



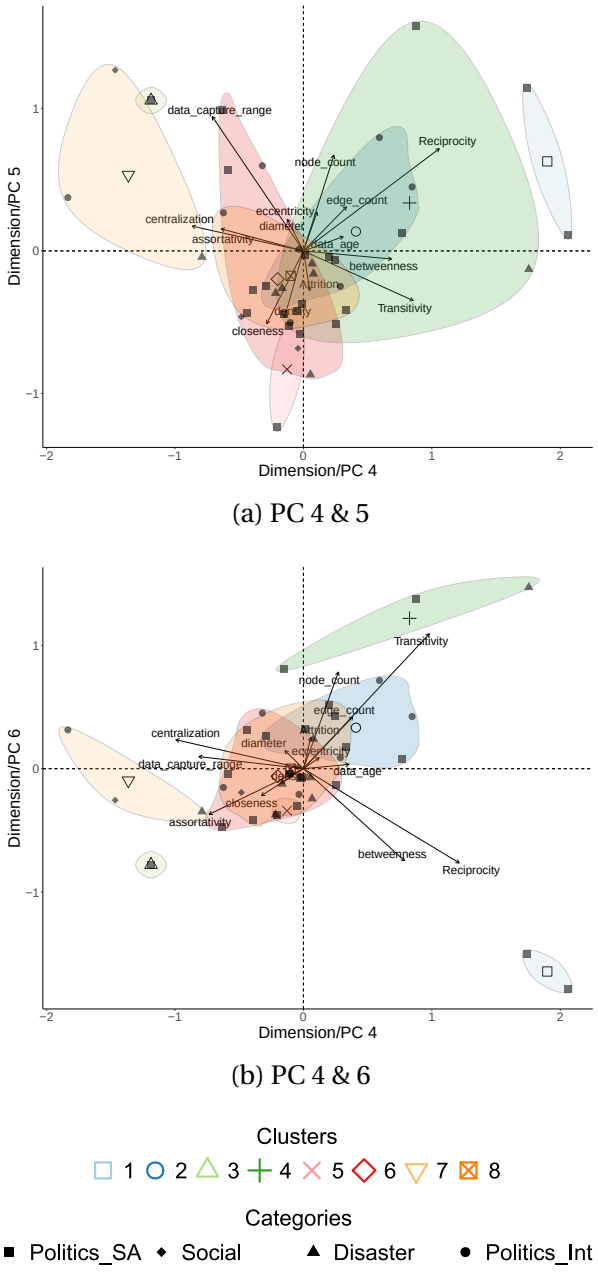
(a) PC 3 & 5

(b) PC 3 & 6

Note: The categories of the data sets visualised on pairs of principal components

Stellenbosch University https://scholar.sun.ac.za

Figure 7: Meta-figure 7 for Cluster Results



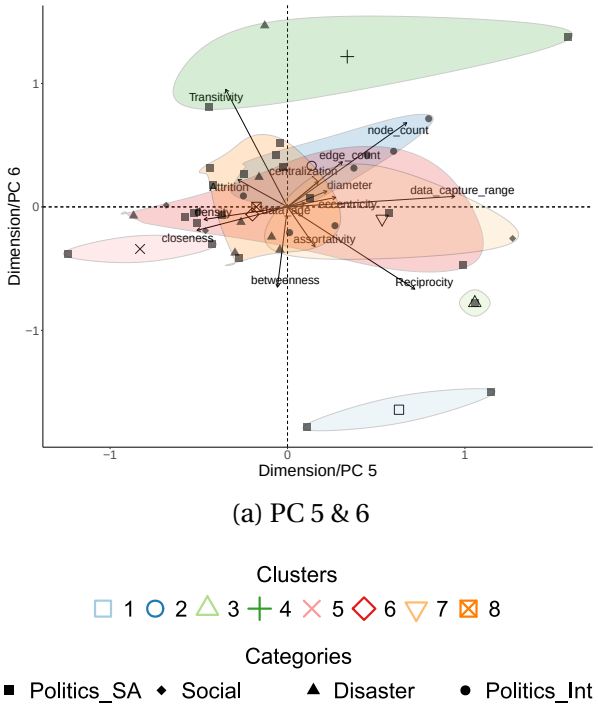
(a) PC 4 & 5

(b) PC 4 & 6

Note: The categories of the data sets visualised on pairs of principal components

Stellenbosch University https://scholar.sun.ac.za

Figure 8: Meta-figure 8 for Cluster Results



(a) PC 5 & 6

Note: The categories of the data sets visualised on pairs of principal components